(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE

(19) World Intellectual Property Organization
International Bureau

EXPRESS MAIL NO.
EV170139843US

(43) International Publication Date
14 June 2001 (14.06.2001)

PCT

(10) International Publication Number
WO 01/42451 A2

(51) International Patent Classification[7]: C12N 15/09, C07K 14/47

(21) International Application Number: PCT/IB00/01938

(22) International Filing Date: 7 December 2000 (07.12.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/169,629	8 December 1999 (08.12.1999)	US
60/187,470	6 March 2000 (06.03.2000)	US

(71) Applicant *(for all designated States except US)*: **GENSET** [FR/FR]; Intellectual Property Department, 24, rue Royale, F-75008 Paris (FR).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DUMAS MILNE EDWARDS, Jean-Baptiste** [FR/FR]; 8, rue Grégoire de Tours, F-75006 Paris (FR). **BOUGUELERET, Lydie** [FR/FR]; 108, avenue Victor Hugo, F-92170 Vanves (FR). **JOBERT, Séverin** [FR/FR]; 7, impasse Tourneux, F-75010 Paris (FR).

(74) Common Representative: **GENSET**; Intellectual Property Department, 24, rue Royale, F-75008 Paris (FR).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/42451 A2

(54) Title: FULL-LENGTH HUMAN cDNAs ENCODING POTENTIALLY SECRETED PROTEINS

(57) Abstract: The invention concerns GENSET polynucleotides and polypeptides. Such GENSET products may be used as reagents in forensic analyses, as chromosome markers, as tissue/cell/organelle-specific markers, in the production of expression vectors. In addition, they may be used in screening and diagnosis assays for abnormal GENSET expression and/or biological activity and for screening compounds that may be used in the treatment of GENSET-related disorders.

Full-length human cDNAs encoding potentially secreted proteins

Related application

The present application claims priority to the US Provisional Patent Applications Serial Nos 60/169,629 and 60/187,470 filed December, 8, 1999, and March, 6, 2000, respectively, the disclosures of which are incorporated herein by reference in their entireties.

Field of the invention

The present invention is directed to polynucleotides encoding GENSET polypeptides, fragments thereof, and the regulatory regions located in the 5'- and 3'-ends of the GENSET genes. The invention also concerns polypeptides encoded by the GENSET polynucleotides and fragments thereof. The present invention also relates to recombinant vectors, which include the polynucleotides of the present invention, particularly recombinant vectors comprising a GENSET regulatory region or a sequence encoding a GENSET polypeptide, and to host cells containing the polynucleotides of the invention, as well as to methods of making such vectors and host cells. The present invention further relates to the use of these recombinant vectors and host cells in the production of the polypeptides of the invention. The invention further relates to antibodies that specifically bind to the polypeptides of the invention and to methods for producing such antibodies and fragments thereof. The invention also provides for methods of detecting the presence of the polynucleotides and polypeptides of the present invention in a sample, methods of diagnosis and screening of abnormal GENSET gene expression and/or biological activity, methods of screening compounds for their ability to modulate the activity or expression of GENSET genes and uses of such compounds.

Background of the invention

The estimated 50,000-100,000 genes scattered along the human chromosomes offer tremendous promise for the understanding, diagnosis, and treatment of human diseases. In addition, probes capable of specifically hybridizing to loci distributed throughout the human genome find applications in the construction of high resolution chromosome maps and in the identification of individuals.

Currently, two different approaches are being pursued for identifying and characterizing the genes distributed along the human genome. In one approach, large fragments of genomic DNA are isolated, cloned, and sequenced. Potential open reading frames in these genomic sequences are identified using bio-informatics software. However, this approach entails sequencing large stretches of human DNA which do not encode proteins in order to find the protein encoding sequences scattered throughout the genome. In addition to requiring extensive sequencing, the bio-

informatics software may mischaracterize the genomic sequences obtained, *i.e.,* labeling non-coding DNA as coding DNA and vice versa.

An alternative approach takes a more direct route to identifying and characterizing human genes. In this approach, complementary DNAs (cDNAs) are synthesized from isolated messenger RNAs (mRNAs) which encode human proteins. Using this approach, sequencing is only performed on DNA which is derived from protein coding fragments of the genome. In the past, these cDNAs, ofter short EST sequences were obtained from oligo-dT primed cDNA libraries. Accordingly, they mainly corresponded to the 3' untranslated region of the mRNA. In part, the prevalence of EST sequences derived from the 3' end of the mRNA is a result of the fact that typical techniques for obtaining cDNAs, are not well suited for isolating cDNA sequences derived from the 5' ends of mRNAs (Adams *et al., Nature* 377:3-174, 1996, Hillier *et al., Genome Res.* 6:807-828, 1996). In addition, in those reported instances where longer cDNA sequences have been obtained, the reported sequences typically correspond to coding sequences and do not include the full 5' untranslated region (5'UTR) of the mRNA from which the cDNA is derived. Indeed, 5'UTRs have been shown to affect either the stability or translation of mRNAs. Thus, regulation of gene expression may be achieved through the use of alternative 5'UTRs as shown, for instance, for the translation of the tissue inhibitor of metalloprotease mRNA in mitogenically activated cells (Waterhouse *et al., J Biol Chem.* 265:5585-9. 1990). Furthermore, modification of 5'UTR through mutation, insertion or translocation events may even be implied in pathogenesis. For instance, the fragile X syndrome, the most common cause of inherited mental retardation, is partly due to an insertion of multiple CGG trinucleotides in the 5'UTR of the fragile X mRNA resulting in the inhibition of protein synthesis via ribosome stalling (Feng *et al., Science* 268:731-4, 1995). An aberrant mutation in regions of the 5'UTR known to inhibit translation of the proto-oncogene *c-myc* was shown to result in upregulation of c-myc protein levels in cells derived from patients with multiple myelomas (Willis *et al., Curr Top Microbiol Immunol* 224:269-76, 1997). In addition, the use of oligo-dT primed cDNA libraries does not allow the isolation of complete 5'UTRs since such incomplete sequences obtained by this process may not include the first exon of the mRNA, particularly in situations where the first exon is short. Furthermore, they may not include some exons, often short ones, which are located upstream of splicing sites. Thus, there is a need to obtain sequences derived from the 5' ends of mRNAs.

Moreover, despite the great amount of EST data that large-scale sequencing projects have yielded (Adams *et al., Nature* 377:174, 1996, Hillier *et al., Genome Res.* 6:807-828, 1996), information concerning the biological function of the mRNAs corresponding to such obtained cDNAs has revealed to be limited. Indeed, whereas the knowledge of the complete coding sequence is absolutely necessary to investigate the biological function of mRNAs, ESTs yield only partial coding sequences. So far, large-scale full-length cDNA cloning has been achieved only with limited success because of the poor efficiency of methods for constructing full-length cDNA

libraries. Indeed, such methods require either a large amount of mRNA (Ederly *et al.*, 1995), thus resulting in non representative full-length libraries when small amounts of tissue are available or require PCR amplification (Maruyama *et al.*, 1994; CLONTECHniques, 1996) to obtain a reasonable number of clones, thus yielding strongly biased cDNA libraries where rare and long cDNAs are lost. Thus, there is a need to obtain full-length cDNAs, *i.e.* cDNAs containing the full coding sequence of their corresponding mRNAs. The present application presents a number of cDNAs, called GENSET polynucleotides, isolated from full-length cDNA librairies obtained from the methods described in PCT publication WO 00/37491.

While many sequences derived from human chromosomes have practical applications, approaches based on the identification and characterization of those chromosomal sequences which encode a protein product are particularly relevant to diagnostic and therapeutic uses. Of the 50,000-100,000 protein coding genes, those genes encoding proteins which are secreted from the cell in which they are synthesized, as well as the secreted proteins themselves, are particularly valuable as potential therapeutic agents. Such proteins are often involved in cell to cell communication and may be responsible for producing a clinically relevant response in their target cells. In fact, several secretory proteins, including tissue plasminogen activator, G-CSF, GM-CSF, erythropoietin, human growth hormone, insulin, interferon-α, interferon-β, interferon-γ, and interleukin-2, are currently in clinical use. These proteins are used to treat a wide range of conditions, including acute myocardial infarction, acute ischemic stroke, anemia, diabetes, growth hormone deficiency, hepatitis, kidney carcinoma, chemotherapy induced neutropenia and multiple sclerosis. For these reasons, cDNAs encoding secreted proteins or fragments thereof represent a particularly valuable source of therapeutic agents. Thus, there is a need for the identification and characterization of secreted proteins and the nucleic acids encoding them.

In addition to being therapeutically useful themselves, secretory proteins include short peptides, called signal peptides, at their amino termini which direct their secretion. These signal peptides are encoded by the signal sequences located at the 5' ends of the coding sequences of genes encoding secreted proteins. Because these signal peptides will direct the extracellular secretion of any protein to which they are operably linked, the signal sequences may be exploited to direct the efficient secretion of any protein by operably linking the signal sequences to a gene encoding the protein for which secretion is desired. In addition, fragments of the signal peptides called membrane-translocating sequences, may also be used to direct the intracellular import of a peptide or protein of interest. This may prove beneficial in gene therapy strategies in which it is desired to deliver a particular gene product to cells other than the cells in which it is produced. Signal sequences encoding signal peptides also find application in simplifying protein purification techniques. In such applications, the extracellular secretion of the desired protein greatly facilitates purification by reducing the number of undesired proteins from which the desired protein must be

selected. Thus, there exists a need to identify and characterize the 5' fragments of the genes for secretory proteins which encode signal peptides.

Sequences coding for human proteins may also find application as therapeutics or diagnostics. In particular, such sequences may be used to determine whether an individual is likely to express a detectable phenotype, such as a disease, as a consequence of a mutation in the coding sequence for a protein. In instances where the individual is at risk of suffering from a disease or other undesirable phenotype as a result of a mutation in such a coding sequence, the undesirable phenotype may be corrected by introducing a normal coding sequence using gene therapy. Alternatively, if the undesirable phenotype results from overexpression of the protein encoded by the coding sequence, expression of the protein may be reduced using antisense or triple helix based strategies.

The GENSET human polypeptides encoded by the coding sequences may also be used as therapeutics by administering them directly to an individual having a condition, such as a disease, resulting from a mutation in the sequence encoding the polypeptide. In such an instance, the condition can be cured or ameliorated by administering the polypeptide to the individual.

In addition, the human polypeptides or fragments thereof may be used to generate antibodies useful in determining the tissue type or species of origin of a biological sample. The antibodies may also be used to determine the subcellular localization of the human polypeptides or the cellular localization of polypeptides which have been fused to the human polypeptides. In addition, the antibodies may also be used in immunoaffinity chromatography techniques to isolate, purify, or enrich the human polypeptide or a target polypeptide which has been fused to the human polypeptide.

Public information on the number of human genes for which the promoters and upstream regulatory regions have been identified and characterized is quite limited. In part, this may be due to the difficulty of isolating such regulatory sequences. Upstream regulatory sequences such as transcription factor binding sites are typically too short to be utilized as probes for isolating promoters from human genomic libraries. Recently, some approaches have been developed to isolate human promoters. One of them consists of making a CpG island library (Cross *et al., Nature Genetics* 6: 236-244, 1994). The second consists of isolating human genomic DNA sequences containing SpeI binding sites by the use of SpeI binding protein. (Mortlock *et al., Genome Res.* 6:327-335, 1996). Both of these approaches have their limits due to a lack of specificity and of comprehensiveness. Thus, there exists a need to identify and systematically characterize the 5' fragments of the genes.

cDNAs including the 5' ends of their corresponding mRNA may be used to efficiently identify and isolate 5'UTRs and upstream regulatory regions which control the location, developmental stage, rate, and quantity of protein synthesis, as well as the stability of the mRNA (Theil *et al., BioFactors* 4:87-93, (1993). Once identified and characterized, these regulatory

regions may be utilized in gene therapy or protein purification schemes to obtain the desired amount and locations of protein synthesis or to inhibit, reduce, or prevent the synthesis of undesirable gene products.

In addition, cDNAs containing the 5' ends of protein genes may include sequences useful as probes for chromosome mapping and the identification of individuals. Thus, there is a need to identify and characterize the sequences upstream of the 5' coding sequences of genes encoding proteins.

Summary of the invention

The present invention provides compositions containing a purified or isolated polynucleotide comprising, consisting of, or consisting essentially of a nucleotide sequence selected from the group consisting of: (a) the sequences of SEQ ID Nos: 1-241; (b) the sequences of clone inserts of the deposited clone pool; (c) the full coding sequences of SEQ ID Nos: 1-241; (d) the full coding sequences of the clone inserts of the deposited clone pool; (e) the sequences encoding one of the polypeptides of SEQ ID Nos: 242-482; (f) the sequences encoding one of the polypeptides encoded by the clone inserts of the deposited clone pool; (g) the genomic sequences coding for GENSET polypeptides; (h) the 5' transcriptional regulatory regions of GENSET genes; (i) the 3' transcriptional regulatory regions of GENSET genes; (j) the polynucleotides comprising the nucleotide sequence of any combination of (g)-(i); (k) the variant polynucleotides of any of the polynucleotides of (a)-(j); (l) the polynucleotides comprising a nucleotide sequence of (a)-(k), wherein the polynucleotide is single stranded, double stranded, or a portion is single stranded and a portion is double stranded; (m) the polynucleotides comprising a nucleotide sequence complementary to any of the single stranded polynucleotides of (l). The invention further provides for fragments of the nucleic acid molecules of (a)-(m) described above.

The present invention also provides biologically active forms, variants, fragments and derivatives of the present proteins, where "biologically active" indicates that the form, variant, fragment, or derivative, has any detectable activity in any in vitro assay known in the art or described herein, or has any detectable function in vivo. In preferred embodiments, a determination of whether a particular polypeptide is biologically active will be made based on any of the specific assays or functional characteristics provided below for each of the proteins of this invention.

Therefore, one embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool, sequences complementary thereto, allelic variants thereof, and degenerate variants thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising at least 8 consecutive nucleotides of a sequence selected from the

group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool, sequences complementary thereto, allelic variants thereof, and degenerate variants thereof. In one aspect of this embodiment, the nucleic acid comprises at least 10, 12, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, 500, 800, 1000, 1500, or 2000 consecutive nucleotides of said selected sequence, sequences complentary thereto, allelic variants thereof, and degenerate variants thereof. The nucleic acid may be a recombinant nucleic acid.

Another embodiment of the present invention is a composition comprising a vertebrate purified or isolated nucleic acid of at least 15,18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, 1000 or 2000 nucleotides in length which hybridizes under stringent conditions to any polynucleotide of the invention, preferably a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool, sequences complementary thereto. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising the full coding sequences of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool, or an allelic variant thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

A further embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising a contiguous span of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted proteins in the deposited clone pool, or an allelic variant thereof, wherein said contiguous span encodes a mature protein. In one aspect of this embodiment, the nucleic acid is recombinant. In another aspect of this embodiment, the nucleic acid is an expression vector wherein said contiguous span which encodes a mature protein is operably linked to a promoter.

Yet another embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising a contiguous span of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted proteins in the deposited clone pool, or an allelic variant thereof, wherein said contiguous span encodes a signal peptide. In one aspect of this embodiment, the nucleic acid is recombinant. In another aspect of this embodiment, the nucleic acid is an fusion vector wherein said contiguous span which encodes a signal peptide is operably linked to a second nucleic acid encoding an heterologous polypeptide.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid encoding a polypeptide comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited

clone pool, or allelic variant thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid encoding a polypeptide comprising the sequence of a mature protein included in a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted proteins in the deposited clone pool, or allelic variant thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid encoding a polypeptide comprising the sequence of a signal peptide included in a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted proteins in the deposited clone pool, or allelic variant thereof. In another aspect it is present in a vector of the invention.

Further embodiments of the invention include compositions containing purified or isolated polynucleotides that comprise, a nucleotide sequence at least 70% identical, more preferably at least 75% identical, and still more preferably at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% identical to any of the polynucleotides of the present invention. Methods of determining identity include those well known in the art and described herein. Such analyses can be performed using a full length polynucleotide sequence or using a subsequence of any length. For example, any two sequences can be compared over a region, in either protein or in both proteins, of any 10, 25, 50, 100, 250, 500, 1000, 2000 or more contiguous nucleotides. In addition, any two sequences can be identified as homologous even when they share sequence homology over a limited region of either polynucleotide, for example over a region of at least about 10, 25, 50, 100, 250, 500, 1000, or more contiguous nucleotides.

The invention further provides compositions containing a purified or isolated polypeptide comprising, consisting of, or consisting essentially of an amino acid sequence selected from the group consisting of: (a) the polypeptides of SEQ ID Nos: 242-482; (b) the polypeptides encoded by the clone inserts of the deposited clone pool; (c) the epitope-bearing fragments of the polypeptides of SEQ ID Nos: 242-482; (d) the epitope-bearing fragments of the polypeptides encoded by the clone inserts contained in the deposited clone pool; (e) the domains of the polypeptides of SEQ ID Nos: 242-482; (f) the domains of the polypeptides encoded by the clone inserts contained in the deposited clone pool; and (g) the allelic variant polypeptides of any of the polypeptides of (a)-(f). The invention further provides for fragments of the polypeptides of (a)-(g) above, such as those having biological activity or comprising biologically functional domain(s).

Yet another embodiment of the present invention is a composition containing a purified or isolated protein comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allelic variant thereof.

Another embodiment of the present invention is a composition containing a purified or isolated polypeptide comprising at least 5, 6 or 8 consecutive amino acids of a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allellic variant thereof. In one aspect of this embodiment, the purified or isolated polypeptide comprises at least 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, 200, 250, 300, 350, 400, 450 or 500 consecutive amino acids of said selected sequence or allelic variant thereof.

Another embodiment of the present invention is a composition containing an isolated or purified polypeptide comprising a signal peptide of a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allellic variant thereof.

Yet another embodiment of the present invention is a composition containing an isolated or purified polypeptide comprising a mature protein of a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allellic variant thereof.

A further embodiment of the present invention are compositions containing polypeptide having an amino acid sequence with at least 70% similarity, and more preferably at least 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% similarity to a polypeptide of the present invention, as well as polypeptides having an amino acid sequence at least 70% identical, more preferably at least 75% identical, and still more preferably 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% identical to a polypeptide of the present invention. Such analyses can be performed using a full length polypeptide sequence or using a subsequence of any length. For example, any two sequences can be compared over a region, in either protein or in both proteins, of any 10, 25, 50, 100, 250, 500, 1000, 2000 or more contiguous amino acids. In addition, any two sequences can be identified as homologous even when they share sequence homology over a limited region of either protein, for example over a region of at least about 10, 25, 50, 100, 250, 500, 1000, or more contiguous amino acids. Further included in the invention are compositions comprising a purified or isolated nucleic acid molecule encoding such polypeptides. Methods for determining identity include those well known in the art and described herein.

The present invention also relates to compositions comprising recombinant vectors, which include the purified or isolated polynucleotides of the present invention, and to host cells recombinant for the polynucleotides of the present invention, as well as to methods of making such vectors and host cells. The present invention further relates to the use of these recombinant vectors and recombinant host cells in the production of GENSET polypeptides.

Consequently, another embodiment of the invention is a vector comprising any polynucleotide of the invention. In a preferred embodiment, the vector is an expression vector comprising a nucleic acid sequence encoding a polypeptide selected from the group consisting of

sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by the clone inserts of the deposited clone pool, or allelic variant thereof, wherein said nucleic acid sequence is operably linked to a promoter. In another preferred embodiment, the vector is a secretion vector comprising a nucleic acid sequence encoding a signal peptide selected from the group consisting of signal peptides of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of secreted polypeptides encoded by the clone inserts of the deposited clone pool, or allelic variant thereof, wherein said nucleic acid sequence is operably linked to an heterologous protein such that said signal peptide will direct the secretion of said heterolgous protein.

A further embodiment of the present invention is a method of making a protein comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, comprising the steps of

a) obtaining a cDNA comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool;

b) inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter; and

c) introducing said expression vector into a host cell whereby the host cell produces the protein encoded by said cDNA.

In one aspect of this embodiment, the method further comprises the step of isolating said protein.

Another embodiment of the present invention is a protein obtainable by the method described in the preceding paragraph.

Another embodiment of the present invention is a method of making a protein comprising the amino acid sequence of the mature protein contained in a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, comprising the steps of

a) obtaining a cDNA comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts of the deposited clone pool, wherein said cDNA encodes a mature protein;

b) inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter; and

c) introducing said expression vector into a host cell whereby the host cell produces the mature protein encoded by said cDNA.

In one aspect of this embodiment, the method further comprises the step of isolating said protein.

Another embodiment of the present invention is a mature protein obtainable by the method described in the preceding paragraph.

Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool or a sequence complementary thereto described herein.

Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising the full coding sequences of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool.

Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising a contiguous span of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts of the deposited clone pool, wherein said contiguous span codes for a mature protein.

Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising a contiguous span of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts of the deposited clone pool, wherein said contiguous span codes for a signal peptide.

The invention further relates to other methods of making the polypeptides of the present invention.

The present invention further relates to transgenic plants or animals, wherein said transgenic plant or animal is transgenic for a polynucleotide of the present invention and expresses a polypeptide of the present invention.

The invention further relates to compositions comprising antibodies that specifically bind to the GENSET polypeptides of the present invention and fragments thereof as well as to methods for producing such antibodies and fragments thereof.

Therefore, another embodiment of the present invention is a composition containing a purified or isolated antibody capable of specifically binding to a protein comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool. In one aspect of this embodiment, the antibody is capable of binding to a polypeptide comprising at least 6 consecutive amino acids, at least 8 consecutive amino acids, or at least 10 consecutive amino acids of said selected sequence.

The invention also provides kits and methods of detecting GENSET gene expression and/or biological activity in a biological sample. One such method involves assaying for the expression of a GENSET polynucleotide in a biological sample using polymerase chain reaction (PCR) to amplify and detect GENSET polynucleotides or Southern and Northern blot hybridization to detect GENSET genomic DNA, cDNA or mRNA. Alternatively, a method of detecting GENSET gene

expression in a test sample can be accomplished using a compound which binds to a GENSET polypeptide of the present invention or a portion of a GENSET polypeptide.

The present invention also relates to diagnostic methods of identifying individuals or non-human animals having elevated or reduced levels of GENSET products, which individuals are likely to benefit from therapies to suppress or enhance GENSET gene expression, respectively and to methods of identifying individuals or non-human animals at increased risk for developing, or present state of having, certain diseases/disorders associated with GENSET gene abnormal expression or biological activity.

The present invention also relates to kits and methods of screening compounds for their ability to modulate (e.g. increase or inhibit) the activity or expression of GENSET genes including compounds that interact with GENSET gene regulatory sequences and compounds that interact directly or indirectly with GENSET polypeptides. Uses of such compounds are also under the scope of the present invention.

The present invention also relates to pharmaceutical or physiologically acceptable compositions comprising, an active agent, the polypeptides, polynucleotides or antibodies of the present invention.

The present invention also relates to computer systems containing cDNA codes and polypeptides codes of sequences of the invention and to computer-related methods of comparing sequences, identifying homology or features using GENSET sequences of the invention.

In another aspect, the present invention provides an isolated polynucleotide, said polynucleotide comprising a nucleic acid sequence encoding i) a polypeptide comprising an amino acid sequence having at least about 80% identity to any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool; or a biologically active fragment of said polypeptide.

In one embodiment, the polypeptide comprises any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of the polypeptides encoded by the clone inserts of the deposited clone pool. In another embodiment, the polypeptide comprises a signal peptide. In another embodiment, the polypeptide is a mature protein. In another embodiment, the nucleic acid sequence has at least about 80% identity over at least about 100 contiguous nucleotides to any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool. In another embodiment, the polynucleotide hybridizes under stringent conditions to a polynucleotide comprising any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool. In another embodiment, the nucleic acid sequence comprises any one of the sequences shown as SEQ ID NOs:1-241 or any one the sequences of the clone inserts of the deposited clone pool. In another embodiment, the polynucleotide is operably linked to a promoter.

In another aspect, the present invention provides an expression vector comprising the polynucleotide operably linked to a promoter. In another aspect, the present invention provides a host cell recombinant for the polynucleotide. In another aspect, the present invention provides a non-human transgenic animal comprising the host cell.

In another aspect, the present invention provides a method of making a GENSET polypeptide, the method comprising a) providing a population of host cells comprising a herein-described polynucleotide and b) culturing the population of host cells under conditions conducive to the production of the polypeptide within said host cells.

In one embodiment, the method further comprises purifying the polypeptide from the population of host cells.

In another aspect, the present invention provides a method of making a GENSET polypeptide, the method comprising a) providing a population of cells comprising a herein-described polynucleotide; b) culturing the population of cells under conditions conducive to the production of the polypeptide within the cells; and c) purifying the polypeptide from the population of cells.

In another aspect, the present invention provides an isolated polynucleotide, the polynucleotide comprising a nucleic acid sequence having at least about 80% identity over at least about 100 contiguous nucleotides to any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

In one embodiment, the polynucleotide hybridizes under stringent conditions to a polynucleotide comprising any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool. In another embodiment, the polynucleotide comprises any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

In another aspect, the present invention provides a biologically active polypeptide encoded by any of the herein-described polynucleotides.

In another aspect, the present invention provides an isolated polypeptide or biologically active fragment thereof, the polypeptide comprising an amino acid sequence having at least about 80% sequence identity to any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool.

In one embodiment, the polypeptide is selectively recognized by an antibody raised against an antigenic polypeptide, or an antigenic fragment thereof, said antigenic polypeptide comprising any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool. In another embodiment, the polypeptide comprises any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool. In another

embodiment, the polypeptide comprises a signal peptide. In another embodiment, the polypeptide is a mature protein.

In another aspect, the present invention provides an antibody that specifically binds to any of ther herein-described polypeptides.

In another aspect, the present invention provides a method of determining whether a GENSET gene is expressed within a mammal, the method comprising the steps of: a) providing a biological sample from said mammal; b) contacting said biological sample with either of: i) a polynucleotide that hybridizes under stringent conditions to the polynucleotide of claim 1; or ii) a polypeptide that specifically binds to the polypeptide of claim 19; and c) detecting the presence or absence of hybridization between the polynucleotide and an RNA species within the sample, or the presence or absence of binding of the polypeptide to a protein within the sample; wherein a detection of the hybridization or of the binding indicates that the GENSET gene is expressed within the mammal.

In one embodiment, the polynucleotide is a primer, and the hybridization is detected by detecting the presence of an amplification product comprising the sequence of the primer. In another embodiment, the polypeptide is an antibody.

In another aspect, the present invention provides a method of determining whether a mammal has an elevated or reduced level of GENSET gene expression, the method comprising the steps of : a) providing a biological sample from the mammal; and b) comparing the amount of any of the herein-described polypeptides, or of an RNA species encoding the polypeptide, within the biological sample with a level detected in or expected from a control sample; wherein an increased amount of the polypeptide or the RNA species within the biological sample compared to the level detected in or expected from the control sample indicates that the mammal has an elevated level of the GENSET gene expression, and wherein a decreased amount of the polypeptide or the RNA species within the biological sample compared to the level detected in or expected from the control sample indicates that the mammal has a reduced level of the GENSET gene expression.

In another aspect, the present invention provides a method of identifying a candidate modulator of a GENSET polypeptide, the method comprising : a) contacting any of the herein-described polypeptides with a test compound; and b) determining whether the compound specifically binds to the polypeptide; wherein a detection that the compound specifically binds to the polypeptide indicates that the compound is a candidate modulator of the GENSET polypeptide.

Brief description of drawings

Figure 1 is a map of the expression vector pPT

Figure 2 is a block diagram of an exemplary computer system.

Figure 3 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the identity levels between the new sequence and the sequences in the database.

Figure 4 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 5 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence.

Brief Description of Tables

Table I provides the applicant's internal designation number assigned to each sequence identification number and indicates whether the sequence is a nucleic acid sequence or a polypeptide sequence, and in which vector the cDNA was cloned.

Table II provides structural features for each cDNA of SEQ ID Nos: 1-241 i.e., the locations of the full coding sequences, the signal peptides, the mature polypeptides, the polyA signal and the polyA site.

Table III lists variants for cDNAs of the present invention.

Table IV provides the positions of fragments which are preferably excluded from the present invention.

Tables Va and b provides the positions of fragments which are preferably excluded or included in the present invention. Table IV and Tables Va, and Table Vb provide for the inclusion and exclusion of polynucleotides independently from each other in addition to those described elsewhere in the specification and is therefore, not meant as limiting description.

Table VI lists known biologically structural and functional domains for the polypeptides of the present invention.

Table VII lists antigenic peaks of predicted antigenic epitopes for polypeptides of the present invention.

Table VIII lists the putative chromosomal location of the polynucleotides of the present invention.

Table IX list the Genset's cDNA libraries of tissues and cell types examined that express the polynucleotides of the present invention.

Table X relates to the bias in spatial distribution of the polynucleotide sequences of the present invention.

Table XI lists predicted subcellular localization for cDNAs of the present invention.

Table XII gives the correspondence between the polynucleotides of the US priority applications, namely the US Provisional Patent Applications Serial Nos 60/169,629 and 60/187, (column entitled "Seq Id No in priority applications") and the polynucleotides of the present application (column entitled "Seq Id No in present application").

Brief description of sequence listing

SEQ ID Nos: 1-31 and 33-143 are the nucleotide sequences of cDNAs encoding a potentially secreted protein. The locations of the ORFs and sequences encoding signal peptides are listed in the accompanying Sequence Listing. In addition, the von Heijne score of the signal peptide computed as described below is listed as the "score" in the accompanying Sequence Listing. The sequence of the signal-peptide is listed as "seq" in the accompanying Sequence Listing. The "/" in the signal peptide sequence indicates the location where proteolytic cleavage of the signal peptide occurs to generate a mature protein. When appropriate, the locations of the first and last nucleotides of the coding sequences, eventually the locations of the first and last nucleotides of the polyA and the locations of the first and last nucleotides of the polyA sites are indicated.

SEQ ID Nos. 32 and 144-241 are the nucleotide sequences of cDNAs in which no sequence encoding a signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a sequence encoding a signal peptide in these nucleic acids. The locations of the ORFs are listed in the accompanying Sequence Listing. When appropriate, the locations of the first and last nucleotides of the coding sequences, eventually the locations of the first and last nucleotides of the polyA and the locations of the first and last nucleotides of the polyA sites are indicated.

SEQ ID Nos: 242-272 and 274-384 are the amino acid sequences of polypeptides which contain a signal peptide. These polypeptides are encoded by the cDNAs of SEQ ID Nos: 1-31 and 33-143 respectively. The location of the signal peptide is listed in the accompanying Sequence Listing.

SEQ ID Nos: 273 and 385-482 are the amino acid sequences of polypeptides in which no signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a signal peptide in these polypeptides. These polypeptides are encoded by the nucleic acids of SEQ ID Nos: 32 and 144-241 respectively.

In accordance with the regulations relating to Sequence Listings, the following codes have been used in the Sequence Listing to describes nucleotide sequences. The code "r" in the sequences indicates that the nucleotide may be a guanine or an adenine. The code "y" in the sequences indicates that the nucleotide may be a thymine or a cytosine. The code "m" in the sequences indicates that the nucleotide may be an adenine or a cytosine. The code "k" in the sequences indicates that the nucleotide may be a guanine or a thymine. The code "s" in the sequences indicates that the nucleotide may be a guanine or a cytosine. The code "w" in the sequences indicates that the nucleotide may be an adenine or an thymine. In addition, all instances of the symbol "n" in the nucleic acid sequences mean that the nucleotide can be adenine, guanine, cytosine or thymine.

In some instances, the polypeptide sequences in the Sequence Listing contain the symbol "Xaa." These "Xaa" symbols indicate either (1) a residue which cannot be identified because of nucleotide sequence ambiguity or (2) a stop codon in the determined sequence where applicants

believe one should not exist (if the sequence were determined more accurately). In some instances, several possible identities of the unknown amino acids may be suggested by the genetic code.

In the case of secreted proteins, it should be noted that, in accordance with the regulations governing Sequence Listings, in the appended Sequence Listing, the encoded protein (i.e. the protein containing the signal peptide and the mature protein or part thereof) extends from an amino acid residue having a negative number through a positively numbered amino acid residue. Thus, the first amino acid of the mature protein resulting from cleavage of the signal peptide is designated as amino acid number 1, and the first amino acid of the signal peptide is designated with the appropriate negative number. However, in the present application, positions on amino acid sequences are always given on the full length polypeptide, the first amino acid of the signal peptide being designated as amino acid number 1.

Detailed description

DEFINITIONS

Before describing the invention in greater detail, the following definitions are set forth to illustrate and define the meaning and scope of the terms used to describe the invention herein.

The terms "GENSET gene", when used herein, encompasses genomic, mRNA and cDNA sequences encoding the GENSET protein, including the 5' and 3' untranslated regions of said sequences.

As used herein, a "secreted" protein is one which, when expressed in a suitable host cell, is transported across or through a membrane, including transport as a result of signal peptides in its amino acid sequence. "Secreted" proteins include without limitation proteins secreted wholly (e.g. soluble proteins), or partially (e.g. receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins which are transported across the membrane of the endoplasmic reticulum. As used herein, a "mature protein" is the polypeptide fragment generated after the cleavage of the signal peptide.

The term "full coding sequence" or open reading frame (ORF) of a GENSET gene, when used herein, refers to the complete coding sequence of said gene. In the case of a secreted protein, the full coding sequence comprises the coding sequence for the signal peptide and the coding sequence for the mature polypeptide. Accordingly, the term "full-length polypeptide" refers to the complete polypeptide encoded by said GENSET gene and in the case of a secreted protein it comprises both the signal peptide and the mature polypeptide. The positions of the full length polypeptides and, in the case of secreted proteins, of signal peptides and mature polypeptides are given in the appended sequence listing.

The term "GENSET biological activity" is intended for polypeptides exhibiting an activity similar, but not necessarily identical, to an activity of the GENSET polypeptide of the invention.

The GENSET biological activity of a given polypeptide may be assessed using a suitable biological assay well known to those skilled in the art such as the one(s) described herein. In contrast, the term "biological activity" refers to any activity that a polypeptide of the invention may have.

The term "corresponding mRNA" refers to the mRNA which was the template for the cDNA synthesis which produced a cDNA of the present invention.

The term "corresponding genomic DNA" refers to the genomic DNA which encodes mRNA which includes the sequence of one of the strands of the cDNA in which thymidine residues in the sequence of the cDNA are replaced by uracil residues in the mRNA.

The term "deposited clone pool" is used herein to refer to the pool of clones entitled GENSET.071PRF deposited in ATCC with the accession number PTA-1218 on January, 21, 2000.

The term "heterologous", when used herein, is intended to designate any polynucleotide or polypeptide other than the GENSET polynucleotide or polypeptide respectively.

The term "isolated" requires that the material be removed from its original environment (e. g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition, and still be isolated in that the vector or composition is not part of its natural environment. For example, a naturally-occurring polynucleotide present in a living animal is not isolated, but the same polynucleotide, separated from some or all of the coexisting materials in the natural system, is isolated. Specifically excluded from the definition of "isolated" are: naturally-occurring chromosomes (such as chromosome spreads), artificial chromosome libraries, genomic libraries, and cDNA libraries that exist either as an *in vitro* nucleic acid preparation or as a transfected/transformed host cell preparation, wherein the host cells are either an *in vitro* heterogeneous preparation or plated as a heterogeneous population of single colonies. Also specifically excluded are the above libraries wherein a specified polynucleotide makes up less than 5% of the number of nucleic acid inserts in the vector molecules. Further specifically excluded are whole cell genomic DNA or whole cell RNA preparations (including said whole cell preparations which are mechanically sheared or enzymatically digested). Further specifically excluded are the above whole cell preparations as either an *in vitro* preparation or as a heterogeneous mixture separated by electrophoresis (including blot transfers of the same) wherein the polynucleotide of the invention has not further been separated from the heterologous polynucleotides in the electrophoresis medium (e.g., further separating by excising a single band from a heterogeneous band population in an agarose gel or nylon blot).

The term "purified" does not require absolute purity; rather, it is intended as a relative definition. Purification of starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly

contemplated. As an example, purification from 0.1 % concentration to 10 % concentration is two orders of magnitude. To illustrate, individual cDNA clones isolated from a cDNA library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The cDNA clones are not naturally occurring as such, but rather are obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The conversion of mRNA into a cDNA library involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the synthetic library by clonal selection. Thus, creating a cDNA library from messenger RNA and subsequently isolating individual clones from that library results in an approximately 10^4-10^6 fold purification of the native message.

The term "purified" is further used herein to describe a polypeptide or polynucleotide of the invention which has been separated from other compounds including, but not limited to, polypeptides or polynucleotides, carbohydrates, lipids, etc. The term "purified" may be used to specify the separation of monomeric polypeptides of the invention from oligomeric forms such as homo- or hetero- dimers, trimers, etc. The term "purified" may also be used to specify the separation of covalently closed polynucleotides from linear polynucleotides. A polynucleotide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polynucleotide sequence and conformation (linear versus covalently close). A substantially pure polypeptide or polynucleotide typically comprises about 50%, preferably 60 to 90% weight/weight of a polypeptide or polynucleotide sample, respectively, more usually about 95%, and preferably is over about 99% pure. Polypeptide and polynucleotide purity, or homogeneity, is indicated by a number of means well known in the art, such as agarose or polyacrylamide gel electrophoresis of a sample, followed by visualizing a single band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art. As an alternative embodiment, purification of the polypeptides and polynucleotides of the present invention may be expressed as "at least" a percent purity relative to heterologous polypeptides and polynucleotides (DNA, RNA or both). As a preferred embodiment, the polypeptides and polynucleotides of the present invention are at least; 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 96%, 98%, 99%, or 100% pure relative to heterologous polypeptides and polynucleotides, respectively. As a further preferred embodiment the polypeptides and polynucleotides have a purity ranging from any number, to the thousandth position, between 90% and 100% (e.g., a polypeptide or polynucleotide at least 99.995% pure) relative to either heterologous polypeptides or polynucleotides, respectively, or as a weight/weight ratio relative to all compounds and molecules other than those existing in the carrier. Each number representing a percent purity, to the thousandth position, may be claimed as individual species of purity.

As used interchangeably herein, the terms "nucleic acid molecule(s)", "oligonucleotide(s)", and "polynucleotide(s)" include RNA or DNA (either single or double stranded, coding,

complementary or antisense), or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form (although each of the above species may be particularly specified). The term "nucleotide" is used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. More precisely, the expression "nucleotide sequence" encompasses the nucleic material itself and is thus not restricted to the sequence information (i.e. the succession of letters chosen among the four base letters) that biochemically characterizes a specific DNA or RNA molecule. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. The term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications such as (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar. For examples of analogous linking groups, purine, pyrimidines, and sugars see for example PCT publication No. WO 95/04064, which disclosure is hereby incorporated by reference in its entirety. Preferred modifications of the present invention include, but are not limited to, 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xantine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid (v) ybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid, 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, and 2,6-diaminopurine. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art. Methylenemethylimino linked oligonucleosides as well as mixed backbone compounds having, may be prepared as described in U.S. Pat. Nos. 5,378,825; 5,386,023; 5,489,677; 5,602,240; and 5,610,289, which disclosures are hereby incorporated by reference in their entireties. Formacetal and thioformacetal linked oligonucleosides may be prepared as described in U.S. Pat. Nos. 5,264,562 and 5,264,564, which disclosures are hereby incorporated by reference in their entireties. Ethylene oxide linked oligonucleosides may be prepared as described in U.S. Pat. No. 5,223,618, which disclosure is hereby incorporated by reference in its entirety. Phosphinate oligonucleotides may be prepared as described in U.S. Pat. No. 5,508,270, which disclosure is hereby incorporated by reference in its entirety. Alkyl phosphonate

oligonucleotides may be prepared as described in U.S. Pat. No. 4,469,863, which disclosure is hereby incorporated by reference in its entirety. 3'-Deoxy-3'-methylene phosphonate oligonucleotides may be prepared as described in U.S. Pat. Nos. 5,610,289 or 5,625,050 which disclosures are hereby incorporated by reference in their entireties. Phosphoramidite oligonucleotides may be prepared as described in U.S. Pat. No. 5,256,775 or U.S. Pat. No. 5,366,878 which disclosures are hereby incorporated by reference in their entireties. Alkylphosphonothioate oligonucleotides may be prepared as described in published PCT applications WO 94/17093 and WO 94/02499 which disclosures are hereby incorporated by reference in their entireties. 3'-Deoxy-3'-amino phosphoramidate oligonucleotides may be prepared as described in U.S. Pat. No. 5,476,925, which disclosure is hereby incorporated by reference in its entirety. Phosphotriester oligonucleotides may be prepared as described in U.S. Pat. No. 5,023,243, which disclosure is hereby incorporated by reference in its entirety. Borano phosphate oligonucleotides may be prepared as described in U.S. Pat. Nos. 5,130,302 and 5,177,198 which disclosures are hereby incorporated by reference in their entireties.

The term "upstream" is used herein to refer to a location which is toward the 5' end of the polynucleotide from a specific reference point.

The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another be virtue of their sequence identities in a manner like that found in double-helical DNA with thymine or uracil residues linked to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, 1995, which disclosure is hereby incorporated by reference in its entirety).

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym from "complementary polynucleotide", "complementary nucleic acid" and "complementary nucleotide sequence". These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind. Unless otherwise stated, all complementary polynucleotides are fully complementary on the whole length of the considered polynucleotide.

The terms "polypeptide" and "protein", used interchangeably herein, refer to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude chemical or post-expression modifications of the polypeptides of the invention, although chemical

or post-expression modifications of these polypeptides may be included excluded as specific embodiments. Therefore, for example, modifications to polypeptides that include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Further, polypeptides with these modifications may be specified as individual species to be included or excluded from the present invention. The natural or other chemical modifications, such as those listed in examples above can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. It will be appreciated that the same type of modification may be present in the same or varying degrees at several sites in a given polypeptide. Also, a given polypeptide may contain many types of modifications. Polypeptides may be branched, for example, as a result of ubiquitination, and they may be cyclic, with or without branching. Modifications include acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphotidylinositol, cross-linking, cyclization, disulfide bond formation, demethylation, formation of covalent cross-links, formation of cysteine, formation of pyroglutamate, formylation, gamma-carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, pegylation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination. (See, for instance Creighton (1993); Seifter *et al.*, (1990); Rattan *et al.*, (1992)). Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems, etc...), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

As used herein, the terms "recombinant polynucleotide" and "polynucleotide construct" are used interchangeably to refer to linear or circular, purified or isolated polynucleotides that have been artificially designed and which comprise at least two nucleotide sequences that are not found as contiguous nucleotide sequences in their initial natural environment. In particular, this terms mean that the polynucleotide or cDNA is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the cDNAs will represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. Backbone molecules according to the present invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid insert of interest. Preferably, the enriched cDNAs represent 15% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. More preferably, the enriched cDNAs represent 50% or more of

the number of nucleic acid inserts in the population of recombinant backbone molecules. In a highly preferred embodiment, the enriched cDNAs represent 90% or more (including any number between 90 and 100%, to the thousandth position, e.g., 99.5%) # of the number of nucleic acid inserts in the population of recombinant backbone molecules.

The term "recombinant polypeptide" is used herein to refer to polypeptides that have been artificially designed and which comprise at least two polypeptide sequences that are not found as contiguous polypeptide sequences in their initial natural environment, or to refer to polypeptides which have been expressed from a recombinant polynucleotide.

As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in a functional relationship. A sequence which is "operably linked" to a regulatory sequence such as a promoter means that said regulatory element is in the correct location and orientation in relation to the nucleic acid to control RNA polymerase initiation and expression of the nucleic acid of interest. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence.

As used herein, the term "non-human animal" refers to any non-human animal, including insects, birds, rodents and more usually mammals. Preferred non-human animals include: primates; farm animals such as swine, goats, sheep, donkeys, cattle, horses, chickens, rabbits; and rodents, preferably rats or mice. As used herein, the term "animal" is used to refer to any species in the animal kingdom, preferably vertebrates, including birds and fish, and more preferable a mammal. Both the terms "animal" and "mammal" expressly embrace human subjects unless preceded with the term "non-human".

The term "domain" refers to an amino acid fragment with specific biological properties. This term encompasses all known structural and linear biological motifs. Examples of such motifs include but are not limited to leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal peptides which direct the secretion of proteins, sites for post-translational modification, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

Although they have distinct meanings, the terms "comprising", "consisting of" and "consisting essentially of" may be interchanged for one another throughout the instant application". The term "having" has the same meaning as "comprising" and may be replaced with either the term "consisting of" or "consisting essentially of".

An "amplification product" refers to a product of any amplification reaction, e.g. PCR, RT-PCR, LCR, etc.

A "modulator" of a protein or other compound refers to any agent that has a functional effect on the protein, including physical binding to the protein, alterations of the quantity or quality of expression of the protein, altering any measurable or detectable activity, property, or behavior of the protein, or in any way interacts with the protein or compound.

"A test compound" can be any molecule that is evaluated for its ability to modulate a protein or other compound.

Unless otherwise specified in the application, nucleotides and amino acids of polynucleotides and polypeptides respectively of the present invention are contiguous and not interrupted by heterologous sequences.

Identity Between Nucleic Acids Or Polypeptides

The terms "percentage of sequence identity" and "percentage homology" are used interchangeably herein to refer to comparisons among polynucleotides and polypeptides, and are determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide or polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Homology is evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTP, FASTA, TFASTA, CLUSTALW, FASTDB (Pearson and Lipman, 1988; Altschul *et al.*, 1990; Thompson *et al.*, 1994; Higgins *et al.*, 1996; Altschul *et al.*, 1990; Altschul *et al.*, 1993; Brutlag *et al*, 1990), the disclosures of which are incorporated by reference in their entireties.

In a particularly preferred embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") which is well known in the art (see, e.g., Karlin and Altschul, 1990; Altschul *et al.*, 1990, 1993, 1997), the disclosures of which are incorporated by reference in their entireties. In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and

(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (i.e., aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet *et al.*, 1992; Henikoff and Henikoff, 1993), the disclosures of which are incorporated by reference in their entireties. Less preferably, the PAM or PAM250 matrices may also be used (see, e.g., Schwartz and Dayhoff, eds., 1978), the disclosure of which is incorporated by reference in its entirety. The BLAST programs evaluate the statistical significance of all high-scoring segment pairs identified, and preferably selects those segments which satisfy a user-specified threshold of significance, such as a user-specified percent homology. Preferably, the statistical significance of a high-scoring segment pair is evaluated using the statistical significance formula of Karlin (see, e.g., Karlin and Altschul, 1990), the disclosure of which is incorporated by reference in its entirety. The BLAST programs may be used with the default parameters or with modified parameters provided by the user.

Another preferred method for determining the best overall match between a query nucleotide sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag *et al.* (1990), the disclosure of which is incorporated by reference in its entirety. In a sequence alignment the query and subject sequences are both DNA sequences. An RNA sequence can be compared by first converting U's to T's. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB alignment of DNA sequences to calculate percent identity are: Matrix=Unitary, k-tuple=4, Mismatch Penalty= 1, Joining Penalty=30, Randomization Group Length=0, Cutoff Score= 1, Gap Penalty=5, Gap Size Penalty 0.05, Window Size=500 or the length of the subject nucleotide sequence, whichever is shorter. If the subject sequence is shorter than the query sequence because of 5' or 3' deletions, not because of internal deletions, a manual correction must be made to the results. This is because the FASTDB program does not account for 5' and 3' truncations of the subject sequence when calculating percent identity. For subject sequences truncated at the 5' or 3'ends, relative to the query sequence, the percent identity is corrected by calculating the number of bases of the query sequence that are 5' and 3' of the subject sequence, which are not matched/aligned, as a percent of the total bases of the query sequence. Whether a nucleotide is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using 10, the specified parameters, to arrive at a final percent identity score. This corrected score is what is used for the purposes of the present invention. Only nucleotides outside the 5' and 3' nucleotides of the subject sequence, as displayed by the FASTDB alignment, which are not matched/aligned with the query sequence, are calculated for the

purposes of manually adjusting the percent identity score. For example, a 90 nucleotide subject sequence is aligned to a 100 nucleotide query sequence to determine percent identity. The deletions occur at the 5' end of the subject sequence and therefore, the FASTDB alignment does not show a matched/alignment of the first 10 nucleotides at 5' end. The 10 unpaired nucleotides represent 10% of the sequence (number of nucleotides at the 5' and 3' ends not matched/total number of nucleotides in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 nucleotides were perfectly matched the final percent identity would be 90%. In another example, a 90 nucleotide subject sequence is compared with a 100 nucleotide query sequence. This time the deletions are internal deletions so that there are no nucleotides on the 5' or 3' of the subject sequence which are not matched/aligned with the query. In this case the percent identity calculated by FASTDB is not manually corrected. Once again, only nucleotides 5' and 3' of the subject sequence which are not matched/aligned with the query sequence are manually corrected. No other manual corrections are made for the purposes of the present invention.

Another preferred method for determining the best overall match between a query amino acid sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag *et al.* (1990). In a sequence alignment the query and subject sequences are both amino acid sequences. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB amino acid alignment are: Matrix=PAM 0, k-tuple=2, Mismatch Penalty= 1, Joining Penalty=20, Randomization Group25Length=0, Cutoff Score= 1, Window Size=sequence length, Gap Penalty=5, Gap Size Penalty=0.05, Window Size=500 or the length of the subject amino acid sequence, whichever is shorter. If the subject sequence is shorter than the query sequence due to N-or C-terminal deletions, not because of internal deletions, the results, in percent identity, must be manually corrected. This is because the FASTDB program does not account for N- and C-terminal truncations of the subject sequence when calculating global percent identity. For subject sequences truncated at the N- and C-termini, relative to the query sequence, the percent identity is corrected by calculating the number of residues of the query sequence that are N- and C- terminal of the subject sequence, which are not matched/aligned with a corresponding subject residue, as a percent of the total bases of the query sequence. Whether a residue is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using the specified parameters, to arrive at a final percent identity score. This final percent identity score is what is used for the purposes of the present invention. Only residues to the N- and C-termini of the subject sequence, which are not matched/aligned with the query sequence, are considered for the purposes of manually adjusting the percent identity score. That is, only query amino acid residues outside the farthest N- and C-terminal residues of the subject sequence. For example, a 90 amino

acid residue subject sequence is aligned with a 100-residue query sequence to determine percent identity. The deletion occurs at the N-terminus of the subject sequence and therefore, the FASTDB alignment does not match/align with the first residues at the N-terminus. The 10 unpaired residues represent 10% of the sequence (number of residues at the N- and C- termini not matched/total number of residues in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 residues were perfectly matched the final percent identity would be 90%. In another example, a 90-residue subject sequence is compared with a 100-residue query sequence. This time the deletions are internal so there are no residues at the N- or C-termini of the subject sequence, which are not matched/aligned with the query. In this case the percent identity calculated by FASTDB is not manually corrected. Once again, only residue positions outside the N- and C-terminal ends of the subject sequence, as displayed in the FASTDB alignment, which are not matched/aligned with the query sequence are manually corrected. No other manual corrections are made for the purposes of the present invention.

The term "percentage of sequence similarity" refers to comparisons between polypeptide sequences and is determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which an identical or equivalent amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence similarity. Similarity is evaluated using any of the variety of sequence comparison algorithms and programs known in the art, including those described above in this section. Equivalent amino acid residues are defined herein in the "Mutated polypeptides" section.

POLYNUCLEOTIDES OF THE INVENTION

The present invention concerns GENSET genomic and cDNA sequences. The present invention encompasses GENSET genes, polynucleotides comprising GENSET genomic and cDNA sequences, as well as fragments and variants thereof. These polynucleotides may be purified, isolated, or recombinant.

Also encompassed by the present invention are allelic variants, orthologs, splice variants, and/or species homologues of the GENSET genes. Procedures known in the art can be used to obtain full-length genes and cDNAs, allelic variants, splice variants, full-length coding portions, orthologs, and/or species homologues of genes and cDNAs corresponding to a nucleotide sequence selected from the group consisting of sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, using information from the sequences disclosed herein or the

clone pool deposited with the ATCC. For example, allelic variants, orthologs and/or species homologues may be isolated and identified by making suitable probes or primers from the sequences provided herein and screening a suitable nucleic acid source for allelic variants and/or the desired homologue using any technique known to those skilled in the art including those described into the section entitled "To find similar sequences".

In a specific embodiment, the polynucleotides of the invention are at least 15, 30, 50, 100, 125, 500, or 1000 continuous nucleotides. In another embodiment, the polynucleotides are less than or equal to 300kb, 200kb, 100kb, 50kb, 10kb, 7.5kb, 5kb, 2.5kb, 2kb, 1.5kb, or 1kb in length. In a further embodiment, polynucleotides of the invention comprise a portion of the coding sequences, as disclosed herein, but do not comprise all or a portion of any intron. In another embodiment, the polynucleotides comprising coding sequences do not contain coding sequences of a genomic flanking gene (i.e., 5' or 3' to the gene of interest in the genome). In other embodiments, the polynucleotides of the invention do not contain the coding sequence of more than 1000, 500, 250, 100, 75, 50, 25, 20, 15, 10, 5, 4, 3, 2, or 1 naturally occurring genomic flanking gene(s).

Deposited clone pool of the invention

Expression of GENSET genes has been shown to lead to the production of at least one mRNA species per GENSET gene, which cDNA sequence is set forth in the appended sequence listing as SEQ ID Nos: 1-241. The cDNAs (SEQ ID Nos: 1-241) corresponding to these GENSET mRNA species were cloned in the vector pBluescriptII SK$^-$ (Stratagene) or one of its derivative called pPT (see figure 1). Cells containing the cloned cDNAs of the present invention are maintained in permanent deposit by the inventors at Genset, S.A., 24 Rue Royale, 75008 Paris, France. Table I provides the applicant's internal designation number (column entitled "Internal designation") assigned to each sequence identification number of SEQ ID Nos: 1-482 (column entitled "Seq Id No") and indicates whether the sequence is a nucleic acid sequence or a polypeptide sequence (column entitled "Type"), and in which vector the cDNA was cloned (column entitled "Vector").

Each cDNA can be removed from the Bluescript vector in which it was inserted by performing a NotI Pst I double digestion to produce the appropriate fragment for each clone provided the cDNA sequence of interest does not contain this restriction site within its sequence. The preferable sites for cDNA removal for those clones inserted into pPT are MunI and HindIII, the sites used for cloning provided the cDNA sequence of interest does not contain this restriction site within its sequence. Alternatively, other restriction enzymes of the multicloning site of the vector may be used to recover the desired insert as indicated by the manufacturer or in figure 1.

Pool of cells containing the cDNAs of the invention, from which the cells containing a particular polynucleotide is obtainable, were also deposited with the American Tissue Culture Collection (ATCC), 10801 University Boulevard, Manassas, VA 20110-2209, United States . Each

cDNA clone has been transfected into separate bacterial cells (E-coli) for these composite deposits. In particular, cells containing the sequences of SEQ ID Nos: 1-241 were deposited on January, 21, 2000 in the pool having ATCC Accession No. PTA-1218 and designated GENSET.071PRF.

Bacterial cells containing a particular clone can be obtained from the composite deposit as follows:

An oligonucleotide probe or probes should be designed to the sequence that is known for that particular clone. This sequence can be derived from the sequences provided herein, or from a combination of those sequences. The design of the oligonucleotide probe should preferably follow these parameters:

(a) It should be designed to an area of the sequence which has the fewest ambiguous bases ("N's"), if any;

(b) Preferably, the probe is designed to have a Tm of approximately 80 degree Celsius (assuming 2 degrees for each A or T and 4 degrees for each G or C). However, probes having melting temperatures between 40 degree Celsius and 80 degree Celsius may also be used provided that specificity is not lost.

The oligonucleotide should preferably be labeled with gamma[^{32}P]ATP (specific activity 6000 Ci/mmole) and T4 polynucleotide kinase using commonly employed techniques for labeling oligonucleotides. Other labeling techniques can also be used. Unincorporated label should preferably be removed by gel filtration chromatography or other established methods. The amount of radioactivity incorporated into the probe should be quantified by measurement in a scintillation counter. Preferably, specific activity of the resulting probe should be approximately $4x10^6$ dpm/pmole.

The bacterial culture containing the pool of full-length clones should preferably be thawed and 100 ul of the stock used to inoculate a sterile culture flask containing 25 ml of sterile L-broth containing ampicillin at 100 ug/ml. The culture should preferably be grown to saturation at 37 degree Celsius, and the saturated culture should preferably be diluted in fresh L-broth. Aliquots of these dilutions should preferably be plated to determine the dilution and volume which will yield approximately 5000 distinct and well-separated colonies on solid bacteriological media containing L-broth containing ampicillin at 100 ug/ml and agar at 1.5% in a 150 mm petri dish when grown overnight at 37 degree Celsius. Other known methods of obtaining distinct, well-separated colonies can also be employed.

Standard colony hybridization procedures should then be used to transfer the colonies to nitrocellulose filters and lyse, denature and bake them.

The filter is then preferably incubated at 65 degree Celsius for 1 hour with gentle agitation in 6X SSC (20X stock is 175.3 g NaCl/liter, 88.2 g Na citrate/liter, adjusted to pH 7.0 with NaOH) containing 0.5% SDS, 100 pg/ml of yeast RNA, and 10 mM EDTA (approximately 10 ml per 150 mm filter). Preferably, the probe is then added to the hybridization mix at a concentration greater

than or equal to $1x10^6$ dpm/ml. The filter is then preferably incubated at 65 degree Celsius with gentle agitation overnight. The filter is then preferably washed in 500 ml of 2X SSC/0.1% SDS at room temperature with gentle shaking for 15 minutes. A third wash with 0.1X SSC/0.5% SDS at 65 degree Celsius for 30 minutes to 1 hour is optional. The filter is then preferably dried and subjected to autoradiography for sufficient time to visualize the positives on the X-ray film. Other known hybridization methods can also be employed.

The positive colonies are picked, grown in culture, and plasmid DNA isolated using standard procedures. The clones can then be verified by restriction analysis, hybridization analysis, or DNA sequencing. The plasmid DNA obtained using these procedures may then be manipulated using standard cloning techniques familiar to those skilled in the art.

Alternatively, to recover cDNA inserts from the pool of bacteria, a PCR can be performed on plasmid DNA isolated using standard procedures and primers designed at both ends of the cDNA insertion, including primers designed in the multicloning site of the vector . For example, a PCR reaction may be conducted using universal primers designed by the plasmid provider or using primers which are specific to the cDNA of interest In the case of Bluescript SK(-), a PCR reaction may be conducted using a primer having the sequence GGAAACAGCTATGACCA and a primer having the sequence GTAAAACGACGGCCAGT. This will produce a DNA fragment including a piece of the multiple cloning site and the cDNA insert. If a specific cDNA of interest is to be recovered, primers may be designed in order to be specific for the 5' end and the 3' end of this cDNA using sequence information available from the appended sequence listing. The PCR product which corresponds to the cDNA of interest can then be manipulated using standard cloning techniques familiar to those skilled in the art.

Therefore, an object of the invention is an isolated, purified, or recombinant polynucleotide comprising a nucleotide sequence selected from the group consisting of cDNA inserts of the deposited clone pool. Moreover, preferred polynucleotides of the invention include purified, isolated, or recombinant GENSET cDNAs consisting of, consisting essentially of, or comprising a nucleotide sequence selected from the group consisting of cDNA inserts of the deposited clone pool.

The polynucleotides of SEQ ID NOs: 1-141 may be interchanged with the corresponding polynucleotides encoded by the human cDNA of the clones inserts of the deposited clone pool. The polypeptides of SEQ ID NOs: 242-482 may be interchanged with the corresponding polypeptides encoded by the human cDNA of the clones inserts of the deposited clone pool. The correspondance between the polynucleotides of SEQ ID Nos: 1-141, the polypeptides of SEQ ID NOs: 242-482 and clones inserts of the deposited clone pool is given in Table I..

cDNA sequences of the invention

Another object of the invention is a purified, isolated, or recombinant polynucleotide comprising a nucleotide sequence selected from the group consisting of sequences of SEQ ID Nos: 1-241, complementary sequences thereto, and fragments thereof. Moreover, preferred polynucleotides of the invention include purified, isolated, or recombinant GENSET cDNAs consisting of, consisting essentially of, or comprising a sequence selected from the group consisting of SEQ ID Nos: 1-241.

Polynucleotides GENSET sequences of SEQ ID Nos: 1-241 were then searched for open reading frames able to encode polypeptides. The GENSET ORFs were also searched to identify potential signal sequence motifs using slight modifications of the procedures disclosed in Von Heijne, *Nucleic Acids Res.* 14:4683-4690, 1986, as described in PCT publication WO 00/37491, the entire disclosures of which are incorporated herein by reference. The GENSET cDNAs of SEQ ID Nos: 1-31 and 33-143 encoding polypeptides of SEQ ID Nos: 242-272 and 274-384 were thus found as containing such signal sequences.

Structural parameters of each of the cDNA of the present invention are described in Table II. Namely, Table II provides, for each cDNA of SEQ ID Nos: 1-241 referred to by its sequence identification number (column entitled "Seq Id No"), the locations of the first and last nucleotides of the coding sequences (listed under the heading "Full Coding Sequence"), and, if applicable, the locations of the signal sequence and the sequence encoding the mature polypeptide in the case of secreted proteins (SEQ ID Nos: 1-31 and 33-143) listed under the headings "Signal Sequence" and "Coding Sequence for the mature Protein" respectively, the locations of the first and last nucleotides of the polyA signals (listed under the heading "Poly A Signal") and the locations of the first and last nucleotides of the polyA sites (listed under the heading "Poly A Site").

Accordingly, the full coding sequence (CDS) or open reading frame (ORF) of each cDNA of the invention refers to the nucleotide sequence beginning with the first nucleotide of the start codon and ending with the last nucleotide of the stop codon (see column entiled "Full coding sequence" of Table II for sequences of Seq Id Nos: 1-241). Similarly, the signal sequence of each cDNA of the invention refers to the nucleotide sequence beginning with the first nucleotide of the start codon and ending with the last nucleotide of the codon encoding the signal peptide (see column entiled "Signal sequence" of Table II for sequences of Seq Id Nos: 1-31 and 33-143) and the coding sequence for the mature polypeptide of each cDNA of the invention refers to the nucleotide sequence beginning with the first nucleotide of the first codon encoding and ending with the last nucleotide of the stop codon (see column entiled "Coding sequence for mature protein" of Table II for sequences of Seq Id Nos: 1-31 and 33-143). Similarly, the 5'untranslated region (or 5'UTR) of each cDNA of the invention refers to the nucleotide sequence starting at nucleotide 1 and ending at the nucleotide immediately 5' to the first nucleotide of the start codon. The 3'untranslated region (or 3'UTR) of each cDNA of the invention refers to the nucleotide sequence

starting at the nucleotide immediately 3' to the last nucleotide of the stop codon and ending at the last nucleotide of the cDNA.

Untranslated regions

In addition, the invention concerns a purified, isolated, and recombinant nucleic acid comprising a nucleotide sequence selected from the group consisting of the 5'UTRs of sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, sequences complementary thereto, and allelic variants thereof. The invention also concerns a purified, isolated, and recombinant nucleic acid comprising a nucleotide sequence selected from the group consisting of the 3'UTRs of sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, sequences complementary thereto, and allelic variants thereof.

These polynucleotides may be used to detect the presence of GENSET mRNA species in a biological sample using either hybridization or RT-PCR techniques well known to those skilled in the art those skilled in the art.

In addition, these polynucleotides may be used as regulatory molecules able to affect the processing and maturation of the polynucleotide including them (either a GENSET polynucleotide or an heterologous polynucleotide), preferably the localization, stability and/or translation of said polynucleotide including them (for a review on UTRs see Decker and Parker, 1995, Derrigo *et al.*, 2000). In particular, 3'UTRs may be used in order to control the stability of heterologous mRNAs in recombinant vectors using any methods known to those skilled in the art including Makrides (1999), US Patents 5,925,56; 5,807,7 and 5,756,264, which disclosures are hereby incorporated by reference in their entireties.

Coding sequences

Another object of the invention is an isolated, purified or recombinant polynucleotide comprising the full coding sequence of a sequence selected from the group consisting of sequences of SEQ ID Nos: 1-241, clone inserts of the deposited clone pool, and variants thereof.

A further object of the invention is an isolated, purified or recombinant polynucleotide encoding a polypeptide comprising a sequence selected from the group consisting of sequences of SEQ ID Nos: 242-482 and allelic variants thereof. Another object of the invention is an isolated, purified or recombinant polynucleotide encoding a polypeptide comprising a sequence selected from the group consisting of polypeptides encoded by cDNA inserts of the deposited clone pool and allelic variants thereof.

In a preferred embodiment, the invention encompasses an isolated, purified or recombinant polynucleotide encoding a polypeptide comprising a sequence selected from the group consisting of the mature proteins of SEQ ID Nos: 242-272 and 274-384. In another preferred embodiment, the invention encompasses an isolated, purified or recombinant polynucleotide encoding a polypeptide

comprising a sequence selected from the group consisting of the signal peptides of SEQ ID Nos: 242-272 and 274-384.

It will be appreciated that should the extent of the full coding sequence differ from that indicated in the appended sequence listing as a result of a sequencing error, reverse transcription or amplification error, mRNA splicing, post-translational modification of the encoded protein, enzymatic cleavage of the encoded protein, or other biological factors, one skilled in the art would be readily able to identify the extent of the full coding sequences in the sequences of SEQ ID Nos: 1-241. Accordingly, the scope of any claims herein relating to nucleic acids containing the full coding sequence of one of SEQ ID Nos: 1-241 is not to be construed as excluding any readily identifiable variations from or equivalents to the full coding sequences described in the appended sequence listing. Similarly, should the extent of the polypeptides differ from those indicated in the appended sequence listing as a result of any of the preceding factors, the scope of claims relating to polypeptides comprising the amino acid sequence of the polypeptides of SEQ ID Nos: 242-482 is not to be construed as excluding any readily identifiable variations from or equivalents to the sequences described in the appended sequence listing.

It will be appreciated that should the extent of the coding sequence of the mature protein differ from that indicated in the appended sequence listing as a result of a sequencing error, reverse transcription or amplification error, mRNA splicing, post-translational modification of the encoded protein, enzymatic cleavage of the encoded protein, or other biological factors, one skilled in the art would be readily able to identify the extent of the coding sequences for the mature protein in the sequences of SEQ ID Nos: 1-31 and 33-143. Accordingly, the scope of any claims herein relating to nucleic acids containing the coding sequence for the mature proteins of one of SEQ ID Nos: 1-31 and 33-143 is not to be construed as excluding any readily identifiable variations from or equivalents to the coding sequences described in the appended sequence listing. Similarly, should the extent of the mature polypeptides differ from those indicated in the appended sequence listing as a result of any of the preceding factors, the scope of claims relating to mature polypeptides comprising the amino acid sequence of the polypeptides of SEQ ID Nos: 242-272 and 274-384 is not to be construed as excluding any readily identifiable variations from or equivalents to the sequences described in the appended sequence listing.

It will be appreciated that should the extent of the coding sequence of the signal peptide differ from that indicated in the appended sequence listing as a result of a sequencing error, reverse transcription or amplification error, mRNA splicing, post-translational modification of the encoded protein, enzymatic cleavage of the encoded protein, or other biological factors, one skilled in the art would be readily able to identify the extent of the coding sequences for the signal peptide in the sequences of SEQ ID Nos: 1-31 and 33-143. Accordingly, the scope of any claims herein relating to nucleic acids containing the signal sequence of one of SEQ ID Nos: 1-31 and 33-143 is not to be construed as excluding any readily identifiable variations from or equivalents to the coding

sequences described in the appended sequence listing. Similarly, should the extent of the signal peptides differ from those indicated in the appended sequence listing as a result of any of the preceding factors, the scope of claims relating to signal peptides comprising the amino acid sequence of the polypeptides of SEQ ID Nos: 242-272 and 274-384 is not to be construed as excluding any readily identifiable variations from or equivalents to the sequences described in the appended sequence listing.

The above disclosed polynucleotides that contains the coding sequence (for the full-length protein of for the mature protein) of the GENSET genes may be expressed in a desired host cell or a desired host organism, when this polynucleotide is placed under the control of suitable expression signals. The expression signals may be either the expression signals contained in the regulatory regions in the GENSET genes of the invention or in contrast the signals may be exogenous regulatory nucleic sequences. Such a polynucleotide, when placed under the suitable expression signals, may also be inserted in a vector for its expression and/or amplification.

Further included in the present invention are polynucleotides encoding the polypeptides of the present invention that are fused in frame to the coding sequences for additional heterologous amino acid sequences. Of special interest are polynucleotides comprising GENSET signal sequences fused to an heterologous polypeptide as described in the section entitled "Secretion vectors". Also included in the present invention are nucleic acids encoding polypeptides of the present invention together with additional, non-coding sequences, including for example, but not limited to non-coding 5' and 3' sequences, vector sequence, sequences used for purification, probing, or priming. For example, heterologous sequences include transcribed, untranslated sequences that may play a role in transcription, and mRNA processing, for example, ribosome binding and stability of mRNA. The heterologous sequences may alternatively comprise additional coding sequences that provide additional functionalities. Thus, a nucleotide sequence encoding a polypeptide may be fused to a tag sequence, such as a sequence encoding a peptide that facilitates purification of the fused polypeptide. In certain preferred embodiments of this aspect of the invention, the tag amino acid sequence is a hexa-histidine peptide, such as the tag provided in a pQE vector (QIAGEN), among others, many of which are commercially available. For instance, hexa-histidine provides for convenient purification of the fusion protein (See Gentz *et al.*, 1989), the disclosure of which is incorporated by reference in its entirety. The "HA" tag is another peptide useful for purification which corresponds to an epitope derived from the influenza hemagglutinin protein (See Wilson *et al.*, 1984), the disclosure of which is incorporated by reference in its entirety. As discussed below other such fusion proteins include the GENSET protein fused to Fc at the N- or C-terminus.

Suitable recombinant vectors that contain a polynucleotide such as described herein are disclosed elsewhere in the specification. Expression vectors encoding GENSET polypeptides or fragments thereof are described in the section entitled "Preparation of the polypeptides".

Regulatory sequences of the invention

As mentioned, the genomic sequence of GENSET genes contains regulatory sequences in the non-coding 5'-flanking region and possibly in the non-coding 3'-flanking region that border the GENSET coding regions containing the exons of these genes.

Polynucleotides derived from GENSET 5' and 3' regulatory regions are useful in order to detect the presence of at least a copy of a genomic nucleotide sequence of the GENSET gene or a fragment thereof in a test sample.

Preferred regulatory sequences

Polynucleotides carrying the regulatory elements located at the 5' end and at the 3' end of GENSET coding regions may be advantageously used to control the transcriptional and translational activity of a heterologous polynucleotide of interest.

Thus, the present invention also concerns a purified or isolated nucleic acid comprising a polynucleotide which is selected from the group consisting of the 5' and 3' GENSET regulatory regions, sequences complementary thereto, regulatory active fragments and variants thereof. The invention also pertains to a purified or isolated nucleic acid comprising a polynucleotide having at least 95% nucleotide identity with a polynucleotide selected from the group consisting of GENSET 5' and 3' regulatory regions, advantageously 99 % nucleotide identity, preferably 99.5% nucleotide identity and most preferably 99.8% nucleotide identity with a polynucleotide selected from the group consisting of GENSET 5' and 3' regulatory regions, sequences complementary thereto, variants and regulatory active fragments thereof.

Another object of the invention consists of purified, isolated or recombinant nucleic acids comprising a polynucleotide that hybridizes, under the stringent hybridization conditions defined herein, with a polynucleotide selected from the group consisting of the nucleotide sequences of GENSET 5'- and 3' regulatory regions, sequences complementary thereto, variants and regulatory active fragments thereof.

Preferred fragments of 5' regulatory regions have a length of about 1500 or 1000 nucleotides, preferably of about 500 nucleotides, more preferably about 400 nucleotides, even more preferably 300 nucleotides and most preferably about 200 nucleotides.

Preferred fragments of 3' regulatory regions are at least 20, 50, 100, 150, 200, 300 or 400 bases in length.

"Providing" with respect to, e.g. a biological sample, population of cells, etc. indicates that the sample, population of cells, etc. is somehow used in a method or procedure. Significantly, "providing" a biological sample or population of cells does not require that the sample or cells are specifically isolated or obtained for the purposes of the invention, but can instead refer, for example, to the use of a biological sample obtained by another individual, for another purpose.

"Regulatory active" polynucleotide derivatives of the 5' regulatory region are polynucleotides comprising or alternatively consisting of a fragment of said polynucleotide which is functional as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide in a recombinant cell host. It could act either as an enhancer or as a repressor. For the purpose of the invention, a nucleic acid or polynucleotide is "functional" as a regulatory region for expressing a recombinant polypeptide or a recombinant polynucleotide if said regulatory polynucleotide contains nucleotide sequences which contain transcriptional and translational regulatory information, and such sequences are "operably linked" to nucleotide sequences which encode the desired polypeptide or the desired polynucleotide.

The regulatory polynucleotides of the invention may be prepared from the nucleotide sequence of GENSET genomic or cDNA sequence, for example, by cleavage using suitable restriction enzymes, or by PCR. The regulatory polynucleotides may also be prepared by digestion of a GENSET gene containing genomic clone by an exonuclease enzyme, such as Bal31 (Wabiko *et al.*, 1986), the disclosure of which is incorporated by reference in its entirety. These regulatory polynucleotides can also be prepared by nucleic acid chemical synthesis, as described elsewhere in the specification.

The regulatory polynucleotides according to the invention may be part of a recombinant expression vector that may be used to express a full coding sequence in a desired host cell or host organism. The recombinant expression vectors according to the invention are described elsewhere in the specification.

Preferred 5'-regulatory polynucleotide of the invention include 5'-UTRs of GENSET cDNAs, or regulatory active fragments or variants thereof. More preferred 5'-regulatory polynucleotides of the invention include sequences selected from the group consisting of 5'-UTRs of sequences of SEQ ID Nos: 1-241, 5'-UTRs of clones inserts of the deposited clone pool, regulatory active fragments and variants thereof.

Preferred 3'-regulatory polynucleotide of the invention include 3'-UTRs of GENSET cDNAs, or regulatory active fragments or variants thereof. More preferred 3'-regulatory polynucleotides of the invention include sequences selected from the group consisting of 3'-UTRs of sequences of SEQ ID Nos: 1-241, 3'-UTRs of clones inserts of the deposited clone pool, regulatory active fragments and variants thereof.

A further object of the invention consists of a purified or isolated nucleic acid comprising:

a) a polynucleotide comprising a 5' regulatory nucleotide sequence selected from the group consisting of:

(i) a nucleotide sequence comprising a polynucleotide of a GENSET 5' regulatory region or a complementary sequence thereto;

(ii) a nucleotide sequence comprising a polynucleotide having at least 95% of nucleotide identity with the nucleotide sequence of a GENSET 5' regulatory region or a complementary sequence thereto;

(iii) a nucleotide sequence comprising a polynucleotide that hybridizes under stringent hybridization conditions with the nucleotide sequence of a GENSET 5' regulatory region or a complementary sequence thereto; and

(iv) a regulatory active fragment or variant of the polynucleotides in (i), (ii) and (iii);

b) a nucleic acid molecule encoding a desired polypeptide or a nucleic acid molecule of interest, said nucleic acid molecule is operably linked to the polynucleotide defined in (a); and

c) optionally, a polynucleotide comprising a 3'- regulatory polynucleotide, preferably a 3'-regulatory polynucleotide of a GENSET gene.

In a specific embodiment, the nucleic acid defined above includes the 5'-UTR of a GENSET cDNA, or a regulatory active fragment or variant thereof.

In a second specific embodiment, the nucleic acid defined above includes the 3'-UTR of a GENSET cDNA, or a regulatory active fragment or variant thereof.

The regulatory polynucleotide of the 5' regulatory region, or its regulatory active fragments or variants, is operably linked at the 5'-end of the nucleic acid molecule encoding the desired polypeptide or nucleic acid molecule of interest.

The regulatory polynucleotide of the 3' regulatory region, or its regulatory active fragments or variants, is advantageously operably linked at the 3'-end of the nucleic acid molecule encoding the desired polypeptide or nucleic acid molecule of interest.

The desired polypeptide encoded by the above-described nucleic acid may be of various nature or origin, encompassing proteins of prokaryotic viral or eukaryotic origin. Among the polypeptides expressed under the control of a GENSET regulatory region include bacterial, fungal or viral antigens. Also encompassed are eukaryotic proteins such as intracellular proteins, such as "house keeping" proteins, membrane-bound proteins, such as mitochondrial membrane-bound proteins and cell surface receptors, and secreted proteins such as endogenous mediators such as cytokines. The desired polypeptide may be an heterologous polypeptide or a GENSET protein, especially a protein with an amino acid sequence selected from the group consisting of sequences of SEQ ID Nos: 242-482, fragments and variants thereof.

The desired nucleic acids encoded by the above-described polynucleotide, usually an RNA molecule, may be complementary to a desired coding polynucleotide, for example to a GENSET coding sequence, and thus useful as an antisense polynucleotide. Such a polynucleotide may be included in a recombinant expression vector in order to express the desired polypeptide or the desired nucleic acid in host cell or in a host organism. Suitable recombinant vectors that contain a polynucleotide such as described herein are disclosed elsewhere in the specification. When a

polynucleotide sequence has been recombinantly introduced into a host cell, the cell is said to be "recombinant" for the polynucleotide.

Polynucleotide variants

The invention also relates to variants of the polynucleotides described herein and fragments thereof. "Variants" of polynucleotides, as the term is used herein, are polynucleotides that differ from a reference polynucleotide. Generally, differences are limited so that the nucleotide sequences of the reference and the variant are closely similar overall and, in many regions, identical. The present invention encompasses both allelic variants and degenerate variants.

Examples of variant sequences of polynucleotides of the invention are given in the appended sequence listing. Table III lists the sequence identification number of all similar sequences of the sequence listing, namely variants. All cDNAS referred to by their sequence identification number on a given line of the table are thought to be variants of the same GENSET gene.

Allelic variant

A variant of a polynucleotide may be a naturally occurring variant such as a naturally occurring allelic variant, or it may be a variant that is not known to occur naturally. By an "allelic variant" is intended one of several alternate forms of a gene occupying a given locus on a chromosome of an organism (see Lewin, 1990), the disclosure of which is incorporated by reference in its entirety. Diploid organisms may be homozygous or heterozygous for an allelic form. Non-naturally occurring variants of the polynucleotide may be made by art-known mutagenesis techniques, including those applied to polynucleotides, cells or organisms.

Degenerate variant

In addition to the isolated polynucleotides of the present invention, and fragments thereof, the invention further includes polynucleotides which comprise a sequence substantially different from those described above but which, due to the degeneracy of the genetic code, still encode a GENSET polypeptide of the present invention. These polynucleotide variants are referred to as "degenerate variants" throughout the instant application. That is, all possible polynucleotide sequences that encode the GENSET polypeptides of the present invention are completed. This includes the genetic code and species-specific codon preferences known in the art. Thus, it would be routine for one skilled in the art to generate the degenerate variants described above, for instance, to optimize codon expression for a particular host (e.g., change codons in the human mRNA to those preferred by other mammalian or bacterial host cells).

Nucleotide changes present in a variant polynucleotide may be silent, which means that they do not alter the amino acids encoded by the polynucleotide. However, nucleotide changes may also result in amino acid substitutions, additions, deletions, fusions and truncations in the

polypeptide encoded by the reference sequence. The substitutions, deletions or additions may involve one or more nucleotides. The variants may be altered in coding or non-coding regions or both. Alterations in the coding regions may produce conservative or non-conservative amino acid substitutions, deletions or additions. In the context of the present invention, preferred embodiments are those in which the polynucleotide variants encode polypeptides which retain substantially the same biological properties or activities as the GENSET protein. More preferred polynucleotide variants are those containing conservative substitutions.

Similar polynucleotides

Other embodiments of the present invention is a purified, isolated or recombinant polynucleotide which is at least 90%, 95%, 96%, 97%, 98% or 99% identical to a polynucleotide selected from the group consisting of sequences of SEQ ID Nos: 1-241 and clone inserts of the deposited clone pool. The above polynucleotides are included regardless of whether they encode a polypeptide having a GENSET biological activity. This is because even where a particular nucleic acid molecule does not encode a polypeptide having activity, one of skill in the art would still know how to use the nucleic acid molecule, for instance, as a hybridization probe or primer. Uses of the nucleic acid molecules of the present invention that do not encode a polypeptide having GENSET activity include, inter alia, isolating a GENSET gene or allelic variants thereof from a DNA library, and detecting GENSET mRNA expression in biological samples, suspected of containing GENSET mRNA or DNA by Northern Blot or PCR analysis.

The present invention is further directed to polynucleotides having sequences at least 50%. 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98% or 99% identity to a polynucleotide selected from the group consisting of sequences of SEQ ID Nos: 1-241 and clone inserts of the deposited clone pool, where said polynucleotide do, in fact, encode a polypeptide having a GENSET biological activity. Of course, due to the degeneracy of the genetic code, one of ordinary skill in the art will immediately recognize that a large number of the polynucleotides at least 50%. 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98%, or 99% identical to a polynucleotide selected from the group consisting of sequences of SEQ ID Nos: 1-241 and clone inserts of the deposited clone pool will encode a polypeptide having biological activity. In fact, since degenerate variants of these nucleotide sequences all encode the same polypeptide, this will be clear to the skilled artisan even without performing the above described comparison assay. It will be further recognized in the art that, for such nucleic acid molecules that are not degenerate variants, a reasonable number will also encode a polypeptide having biological activity. This is because the skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly effect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid), as further described below. By a polynucleotide having a nucleotide sequence at least, for example, 95% "identical" to a reference nucleotide sequence of the present invention, it is intended that the

nucleotide sequence of the polynucleotide is identical to the reference sequence except that the polynucleotide sequence may include up to five point mutations per each 100 nucleotides of the reference nucleotide sequence encoding the GENSET polypeptide. In other words, to obtain a polynucleotide having a nucleotide sequence at least 95% identical to a reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted, inserted, or substituted with another nucleotide. The query sequence may be an entire sequence selected from the group consisting of sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, or the ORF (open reading frame) of a polynucleotide sequence selected from said group, or any fragment specified as described herein.

Hybridizing Polynucleotides

In another aspect, the invention provides an isolated or purified nucleic acid molecule comprising a polynucleotide which hybridizes under stringent hybridization conditions to any polynucleotide of the present invention using any methods known to those skilled in the art including those disclosed herein and in particular in the "To find similar sequences" section. Also contemplated are nucleic acid molecules that hybridize to the polynucleotides of the present invention at lower stringency hybridization conditions, preferably at moderate or low stringency conditions as defined herein. Such hybridizing polynucleotides may be of at least 15,18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, 1000 or 2000 nucleotides in length.

Of particular interest, are the polynucleotides hybridizing to any polynucleotide of the invention and encoding GENSET polypeptides, particularly GENSET polypeptides exhibiting a GENSET biological activity.

Of course, a polynucleotide which hybridizes only to polyA+ sequences (such as any 3' terminal polyA+ tract of a cDNA shown in the sequence listing), or to a 5' complementary stretch of T (or U) residues, would not be included in the definition of "polynucleotide," since such a polynucleotide would hybridize to any nucleic acid molecule containing a poly (A) stretch or the complement thereof (e.g., practically any double-stranded cDNA clone generated using oligo dT as a primer).

Complementary polynucleotides

The invention further provides isolated nucleic acid molecules having a nucleotide sequence fully complementary to any polynucleotide of the invention. The present invention encompasses a purified, isolated or recombinant polynucleotide having a nucleotide sequence complementary to a sequence selected from the group consisting of sequences of SEQ ID Nos: 1-241, sequences of clone inserts of the deposited clone pool and fragments thereof. Such isolated molecules, particularly DNA molecules, are useful as probes for gene mapping and for identifying GENSET mRNA in a biological sample, for instance, by PCR or Northern blot analysis.

Polynucleotides fragments

The present invention is further directed to polynucleotides encoding portions or fragments of the nucleotide sequences described herein. Uses for the polynucleotide fragments of the present invention include probes, primers, molecular weight markers and for expressing the polypeptide fragments of the present invention. Fragments include portions of polynucleotides selected from the group consisting of a) the sequences of SEQ ID Nos: 1-241, b) the genomic GENSET sequences, c) the polynucleotides encoding a polypeptide selected from the group consisting of the sequences of SEQ ID Nos: 242-482, d) the sequences of clone inserts of the deposited clone pool, and e) the polynucleotides encoding the polypeptides encoded by the clone inserts of the deposited clone pool. Particularly included in the present invention is a purified or isolated polynucleotide comprising at least 8 consecutive bases of a polynucleotide of the present invention. In one aspect of this embodiment, the polynucleotide comprises at least 10, 12, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, 500, 800, 1000, 1500, or 2000 consecutive nucleotides of a polynucleotide of the present invention.

In addition to the above preferred polynucleotide sizes, further preferred sub-genuses of polynucleotides comprise at least 8 nucleotides, wherein "at least 8" is defined as any integer between 8 and the integer representing the 3' most nucleotide position as set forth in the sequence listing or elsewhere herein. Further included as preferred polynucleotides of the present invention are polynucleotide fragments at least 8 nucleotides in length, as described above, that are further specified in terms of their 5' and 3' position. The 5' and 3' positions are represented by the position numbers set forth in the appended sequence listing. For allelic, degenerate and other variants, position 1 is defined as the 5' most nucleotide of the ORF, i.e., the nucleotide "A" of the start codon with the remaining nucleotides numbered consecutively. Therefore, every combination of a 5' and 3' nucleotide position that a polynucleotide fragment of the present invention, at least 8 contiguous nucleotides in length, could occupy on a polynucleotide of the invention is included in the invention as an individual species. The polynucleotide fragments specified by 5' and 3' positions can be immediately envisaged and are therefore not individually listed solely for the purpose of not unnecessarily lengthening the specifications.

It is noted that the above species of polynucleotide fragments of the present invention may alternatively be described by the formula "a to b"; where "a" equals the 5' most nucleotide position and "b" equals the 3' most nucleotide position of the polynucleotide; and further where "a" equals an integer between 1 and the number of nucleotides of the polynucleotide sequence of the present invention minus 8, and where "b" equals an integer between 9 and the number of nucleotides of the polynucleotide sequence of the present invention; and where "a" is an integer smaller then "b" by at least 8.

The present invention also provides for the exclusion of any species of polynucleotide fragments of the present invention specified by 5' and 3' positions or sub-genuses of

polynucleotides specified by size in nucleotides as described above. Any number of fragments specified by 5' and 3' positions or by size in nucleotides, as described above, may be excluded. Specifically excluded from the invention are the fragments described in Table IV. For these cDNAs referred to by their sequence identification numbers, Table IV gives the positions of excluded fragments within these sequences fragments having substantial homology to polyadenylation tails and to repeated sequences including Alu, L1, THE and MER repeats, SSTR sequences or satellite, micro-satellite, and telomeric repeats. Each fragment is represented by a-b where a and b are the start and end positions respectively of a given excluded fragment. Excluded fragments are separated from each other by a coma. As used herein the term "polynucleotide described in Table IV" refers to all polynucleotide fragments defined in Table IV in this manner.

Preferred included and excluded polynucleotide fragments of the invention are also described in Tables Va and Table Vb. For these cDNAs referred to by their sequence identification numbers, Tables Va and Table Vb give the positions of preferred fragments within these sequences (columns entitled "Preferentially included fragments") as well as the positions of preferentially excluded fragments (columns entitled "Preferentially excluded fragments"). Each fragment is represented by a-b where a and b are the start and end positions respectively of a given preferred fragment. Fragments are separated from each other by a coma. As used herein the term "excluded polynucleotide described in Tables Va and Vb" refers to all polynucleotide preferentially excluded as described in Tables Va and Vb. As used herein the term "preferred polynucleotide described in Tables Va and Vb" refers to all preferrentially included polynucleotide fragments listed in Tables Va and Table Vb in this manner.

Therefore, the present invention encompasses isolated, purified, or recombinant polynucleotides which consist of, consist essentially of, or comprise a contiguous span of at least 8, 10, 12, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, 500, 1000 or 2000 nucleotides of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and sequences fully complementary thereto, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, wherein said contiguous span comprises at least 1, 2, 3, 5, 10, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400 or 500 nucleotides of a preferred polynucleotide described in Tables Va and Vb, or a sequence complementary thereto. The present invention also encompasses isolated, purified, or recombinant polynucleotides comprising, consisting essentially of, or consisting of a contiguous span of at least 8, 10, 12, 15, 18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, 500, 1000 or 2000 nucleotides of a polynucleotide selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and sequences fully complementary thereto, wherein said contiguous span comprises a preferred polynucleotide described in Tables Va and Vb, or a sequence complementary thereto, to the extent that a contiguous span of these lengths is consistent with the length of the selected sequence. The present invention also encompasses isolated, purified, or recombinant nucleic acids which comprise,

consist of or consist essentially of a contiguous span of a polynucleotide selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and sequences fully complementary thereto, wherein said contiguous span comprises preferred polynucleotide described in Tables Va and Vb, or a sequence complementary thereto.

Other preferred fragments of the invention are polynucleotides comprising polynucleotides encoding domains of polypeptides. Such fragments may be used to obtain other polynucleotides encoding polypeptides having similar domains using hybridization or RT-PCR techniques. Alternatively, these fragments may be used to express a polypeptide domain which may present a specific biological property. Preferred domains for the GENSET polypeptides of the invention are described in Table VI. Thus, another object of the invention is an isolated, purified or recombinant polynucleotide encoding a polypeptide consisting of, consisting essentially of, or comprising a contiguous span of at least 5, 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, 200, 250, 300, 350, 400, 450 or 500 consecutive amino acids of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, where said contiguous span comprises at least 1, 2, 3, 5, or 10 of the amino acid positions of a domain of said selected sequence. The present invention also encompasses isolated, purified or recombinant polynucleotides encoding a polypeptide comprising a contiguous span of at least 5, 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, 200, 250, 300, 350, 400, 450 or 500 consecutive amino acids of a sequence selected from the group consisting of sequences of SEQ ID Nos: 242-482, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, where said contiguous span is a domain of said selected sequence. The present invention also encompasses isolated, purified or recombinant polynucleotides encoding a polypeptide comprising a domain of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482.

The present invention further encompasses any combination of the polynucleotide fragments listed in this section.

Oligonucleotide primers and probes

The present invention also encompasses fragments of GENSET polynucleotides for use as primers and probes. Polynucleotides derived from the GENSET genomic and cDNA sequences are useful in order to detect the presence of at least a copy of a GENSET polynucleotide or fragment, complement, or variant thereof in a test sample.

Structural definition

Any polynucleotide of the invention may be used as a primer or probe. Particularly preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000, 1500 or 2000 nucleotides of a sequence selected from the group

consisting of the GENSET genomic sequences, the cDNA sequences and the sequences fully complementary thereto. Another object of the invention is a purified, isolated, or recombinant polynucleotide comprising the nucleotide sequence of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 1-241, sequences of clone inserts of the deposited clone pool, sequences fully complementary thereto, allelic variants thereof, and fragments thereof. Moreover, preferred probes and primers of the invention include purified, isolated, or recombinant GENSET cDNAs consisting of, consisting essentially of, or comprising the sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool. Particularly preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000, 1500 or 2000 nucleotides of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and the sequences fully complementary thereto.

Design of primers and probes

A probe or a primer according to the invention has between 8 and 1000 nucleotides in length, or is specified to be at least 12, 15, 18, 20, 25, 35, 40, 50, 60, 70, 80, 100, 250, 500, 1000, 1500 or 2000 nucleotides in length. More particularly, the length of these probes and primers can range from 8, 10, 15, 20, or 30 to 100 nucleotides, preferably from 10 to 50, more preferably from 15 to 30 nucleotides. Shorter probes and primers tend to lack specificity for a target nucleic acid sequence and generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. Longer probes and primers are expensive to produce and can sometimes self-hybridize to form hairpin structures. The appropriate length for primers and probes under a particular set of assay conditions may be empirically determined by one of skill in the art. The formation of stable hybrids depends on the melting temperature (Tm) of the DNA. The Tm depends on the length of the primer or probe, the ionic strength of the solution and the G+C content. The higher the G+C content of the primer or probe, the higher is the melting temperature because G:C pairs are held by three H bonds whereas A:T pairs have only two. The GC content in the probes of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

For amplification purposes, pairs of primers with approximately the same Tm are preferable. Primers may be designed using the OSP software (Hillier and Green, 1991), the disclosure of which is incorporated by reference in its entirety, based on GC content and melting temperatures of oligonucleotides, or using PC-Rare (http://bioinformatics.weizmann.ac.il/software/PC-Rare/doc/manuel.html) based on the octamer frequency disparity method (Griffais *et al.*, 1991), the disclosure of which is incorporated by reference in its entirety. DNA amplification techniques are well known to those skilled in the art. Amplification techniques that can be used in the context of the present invention include, but are not limited to, the ligase chain reaction (LCR) described in EP-A- 320 308, WO 9320227 and EP-A-439 182, the

polymerase chain reaction (PCR, RT-PCR) and techniques such as the nucleic acid sequence based amplification (NASBA) described in Guatelli *et al.*(1990) and in Compton (1991), Q-beta amplification as described in European Patent Application No 4544610, strand displacement amplification as described in Walker *et al.* (1996) and EP A 684 315 and, target mediated amplification as described in PCT Publication WO 9322461, the disclosures of which are incorporated by reference in their entireties.

LCR and Gap LCR are exponential amplification techniques, both depend on DNA ligase to join adjacent primers annealed to a DNA molecule. In Ligase Chain Reaction (LCR), probe pairs are used which include two primary (first and second) and two secondary (third and fourth) probes, all of which are employed in molar excess to target. The first probe hybridizes to a first segment of the target strand and the second probe hybridizes to a second segment of the target strand, the first and second segments being contiguous so that the primary probes abut one another in 5' phosphate-3'hydroxyl relationship, and so that a ligase can covalently fuse or ligate the two probes into a fused product. In addition, a third (secondary) probe can hybridize to a portion of the first probe and a fourth (secondary) probe can hybridize to a portion of the second probe in a similar abutting fashion. Of course, if the target is initially double stranded, the secondary probes also will hybridize to the target complement in the first instance. Once the ligated strand of primary probes is separated from the target strand, it will hybridize with the third and fourth probes, which can be ligated to form a complementary, secondary ligated product. It is important to realize that the ligated products are functionally equivalent to either the target or its complement. By repeated cycles of hybridization and ligation, amplification of the target sequence is achieved. A method for multiplex LCR has also been described (WO 9320227), the disclosure of which is incorporated by reference in its entirety. Gap LCR (GLCR) is a version of LCR where the probes are not adjacent but are separated by 2 to 3 bases.

For amplification of mRNAs, it is within the scope of the present invention to reverse transcribe mRNA into cDNA followed by polymerase chain reaction (RT-PCR); or, to use a single enzyme for both steps as described in U.S. Patent No. 5,322,770 or, to use Asymmetric Gap LCR (RT-AGLCR) as described by Marshall *et al.*(1994), the disclosures of which are incorporated by reference in its entireties. AGLCR is a modification of GLCR that allows the amplification of RNA.

The PCR technology is the preferred amplification technique used in the present invention. A variety of PCR techniques are familiar to those skilled in the art. For a review of PCR technology, see White (1997), Erlich (1992) and the publication entitled "PCR Methods and Applications" (1991, Cold Spring Harbor Laboratory Press), the disclosures of which are incorporated by reference in its entireties. In each of these PCR procedures, PCR primers on either side of the nucleic acid sequences to be amplified are added to a suitably prepared nucleic acid sample along with dNTPs and a thermostable polymerase such as Taq polymerase, Pfu polymerase,

Tth polymerase or Vent polymerase. The nucleic acid in the sample is denatured and the PCR primers are specifically hybridized to complementary nucleic acid sequences in the sample. The hybridized primers are extended. Thereafter, another cycle of denaturation, hybridization, and extension is initiated. The cycles are repeated multiple times to produce an amplified fragment containing the nucleic acid sequence between the primer sites. PCR has further been described in several patents including US Patents 4,683,195; 4,683,202; and 4,965,188, the disclosures of which are incorporated herein by reference in their entireties.

Preparation of primers and probes

The primers and probes can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphodiester method of Narang *et al.*(1979), the phosphodiester method of Brown *et al.*(1979), the diethylphosphoramidite method of Beaucage *et al.*(1981) and the solid support method described in EP 0 707 592, which disclosures are hereby incorporated by reference in their entireties.

Detection probes are generally nucleic acid sequences or uncharged nucleic acid analogs such as, for example peptide nucleic acids which are disclosed in International Patent Application WO 92/20702, morpholino analogs which are described in U.S. Patents Numbered 5,185,444; 5,034,506 and 5,142,047, which disclosures are hereby incorporated by reference in their entireties. The probe may have to be rendered "non-extendable" in that additional dNTPs cannot be added to the probe. In and of themselves analogs usually are non-extendable and nucleic acid probes can be rendered non-extendable by modifying the 3' end of the probe such that the hydroxyl group is no longer capable of participating in elongation. For example, the 3' end of the probe can be functionalized with the capture or detection label to thereby consume or otherwise block the hydroxyl group. Alternatively, the 3' hydroxyl group simply can be cleaved, replaced or modified, U.S. Patent Application Serial No. 07/049,061 filed April 19, 1993, which disclosure is hereby incorporated by reference in its entirety, describes modifications, which can be used to render a probe non-extendable.

Labeling of probes

Any of the polynucleotides of the present invention can be labeled, if desired, by incorporating any label known in the art to be detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include radioactive substances (including, ^{32}P, ^{35}S, ^{3}H, ^{125}I), fluorescent dyes (including, 5-bromodesoxyuridin, fluorescein, acetylaminofluorene, digoxigenin) or biotin. Preferably, polynucleotides are labeled at their 3' and 5' ends. Examples of non-radioactive labeling of nucleic acid fragments are described in the French patent No. FR-7810975 or by Urdea *et al* (1988) or Sanchez-Pescador *et al* (1988), which disclosures are hereby incorporated by reference in their entireties. In addition, the probes

according to the present invention may have structural characteristics such that they allow the signal amplification, such structural characteristics being, for example, branched DNA probes as those described by Urdea *et al.* in 1991 or in the European patent No. EP 0 225 807 (Chiron), which disclosures are hereby incorporated by reference in their entireties.

The detectable probe may be single stranded or double stranded and may be made using techniques known in the art, including *in vitro* transcription, nick translation, or kinase reactions. A nucleic acid sample containing a sequence capable of hybridizing to the labeled probe is contacted with the labeled probe. If the nucleic acid in the sample is double stranded, it may be denatured prior to contacting the probe. In some applications, the nucleic acid sample may be immobilized on a surface such as a nitrocellulose or nylon membrane. The nucleic acid sample may comprise nucleic acids obtained from a variety of sources, including genomic DNA, cDNA libraries, RNA, or tissue samples.

Procedures used to detect the presence of nucleic acids capable of hybridizing to the detectable probe include well known techniques such as Southern blotting, Northern blotting, dot blotting, colony hybridization, and plaque hybridization. In some applications, the nucleic acid capable of hybridizing to the labeled probe may be cloned into vectors such as expression vectors, sequencing vectors, or *in vitro* transcription vectors to facilitate the characterization and expression of the hybridizing nucleic acids in the sample. For example, such techniques may be used to isolate and clone sequences in a genomic library or cDNA library which are capable of hybridizing to the detectable probe as described herein.

Immobilization of probes

A label can also be used to capture the primer, so as to facilitate the immobilization of either the primer or a primer extension product, such as amplified DNA, on a solid support. A capture label is attached to the primers or probes and can be a specific binding member which forms a binding pair with the solid's phase reagent's specific binding member (e.g. biotin and streptavidin). Therefore depending upon the type of label carried by a polynucleotide or a probe, it may be employed to capture or to detect the target DNA. Further, it will be understood that the polynucleotides, primers or probes provided herein, may, themselves, serve as the capture label. For example, in the case where a solid phase reagent's binding member is a nucleic acid sequence, it may be selected such that it binds a complementary portion of a primer or probe to thereby immobilize the primer or probe to the solid phase. In cases where a polynucleotide probe itself serves as the binding member, those skilled in the art will recognize that the probe will contain a sequence or "tail" that is not complementary to the target. In the case where a polynucleotide primer itself serves as the capture label, at least a portion of the primer will be free to hybridize with a nucleic acid on a solid phase. DNA Labeling techniques are well known to the skilled technician.

The probes of the present invention are useful for a number of purposes. They can be notably used in Southern hybridization to genomic DNA. The probes can also be used to detect PCR amplification products. They may also be used to detect mismatches in the GENSET gene or mRNA using other techniques.

Any of the polynucleotides, primers and probes of the present invention can be conveniently immobilized on a solid support. The solid support is not critical and can be selected by one skilled in the art. Thus, latex particles, microparticles, magnetic beads, non-magnetic beads (including polystyrene beads), membranes (including nitrocellulose strips), plastic tubes, walls of microtiter wells, glass or silicon chips, sheep (or other suitable animal's) red blood cells and duracytes are all suitable examples. Suitable methods for immobilizing nucleic acids on solid phases include ionic, hydrophobic, covalent interactions and the like. A solid support, as used herein, refers to any material which is insoluble, or can be made insoluble by a subsequent reaction. The solid support can be chosen for its intrinsic ability to attract and immobilize the capture reagent. Alternatively, the solid phase can retain an additional receptor which has the ability to attract and immobilize the capture reagent. The additional receptor can include a charged substance that is oppositely charged with respect to the capture reagent itself or to a charged substance conjugated to the capture reagent. As yet another alternative, the receptor molecule can be any specific binding member which is immobilized upon (attached to) the solid support and which has the ability to immobilize the capture reagent through a specific binding reaction. The receptor molecule enables the indirect binding of the capture reagent to a solid support material before the performance of the assay or during the performance of the assay. The solid phase thus can be a plastic, derivatized plastic, magnetic or non-magnetic metal, glass or silicon surface of a test tube, microtiter well, sheet, bead, microparticle, chip, sheep (or other suitable animal's) red blood cells, duracytes® and other configurations known to those of ordinary skill in the art. The polynucleotides of the invention can be attached to or immobilized on a solid support individually or in groups of at least 2, 5, 8, 10, 12, 15, 20, or 25 distinct polynucleotides of the invention to a single solid support. In addition, polynucleotides other than those of the invention may be attached to the same solid support as one or more polynucleotides of the invention.

Oligonucleotide array

A substrate comprising a plurality of oligonucleotide primers or probes of the invention may be used either for detecting or amplifying targeted sequences in GENSET genes, may also be used for detecting mutations in the coding or in the non-coding sequences of GENSET genes, and may also be used to determine GENSET gene expression in different contexts such as in different tissues, at different stages of a process (embryo development, disease treatment), and in patients versus healthy individuals as described elsewhere in the application.

As used herein, the term "array" means a one dimensional, two dimensional, or multidimensional arrangement of nucleic acids of sufficient length to permit specific detection of gene expression. For example, the array may contain a plurality of nucleic acids derived from genes whose expression levels are to be assessed. The array may include a GENSET genomic DNA, a GENSET cDNA, sequences complementary thereto or fragments thereof. Preferably, the fragments are at least 12, 15, 18, 20, 25, 30, 35, 40 or 50 nucleotides in length. More preferably, the fragments are at least 100 nucleotides in length. Even more preferably, the fragments are more than 100 nucleotides in length. In some embodiments the fragments may be more than 500 nucleotides in length.

Any polynucleotide provided herein may be attached in overlapping areas or at random locations on the solid support. Alternatively the polynucleotides of the invention may be attached in an ordered array wherein each polynucleotide is attached to a distinct region of the solid support which does not overlap with the attachment site of any other polynucleotide. Preferably, such an ordered array of polynucleotides is designed to be "addressable" where the distinct locations are recorded and can be accessed as part of an assay procedure. Addressable polynucleotide arrays typically comprise a plurality of different oligonucleotide probes that are coupled to a surface of a substrate in different known locations. The knowledge of the precise location of each polynucleotides location makes these "addressable" arrays particularly useful in hybridization assays. Any addressable array technology known in the art can be employed with the polynucleotides of the invention. One particular embodiment of these polynucleotide arrays is known as the Genechips™, and has been generally described in US Patent 5,143,854; PCT publications WO 90/15070 and 92/10092, which disclosures are hereby incorporated by reference in their entireties. These arrays may generally be produced using mechanical synthesis methods or light directed synthesis methods which incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis (Fodor *et al.*, 1991), which disclosure is hereby incorporated by reference in its entirety. The immobilization of arrays of oligonucleotides on solid supports has been rendered possible by the development of a technology generally identified as "Very Large Scale Immobilized Polymer Synthesis" (VLSIPS™) in which, typically, probes are immobilized in a high density array on a solid surface of a chip. Examples of VLSIPS™ technologies are provided in US Patents 5,143,854; and 5,412,087 and in PCT Publications WO 90/15070, WO 92/10092 and WO 95/11995, which disclosures are hereby incorporated by reference in their entireties, which describe methods for forming oligonucleotide arrays through techniques such as light-directed synthesis techniques. In designing strategies aimed at providing arrays of nucleotides immobilized on solid supports, further presentation strategies were developed to order and display the oligonucleotide arrays on the chips in an attempt to maximize hybridization patterns and sequence information. Examples of such presentation strategies are disclosed in PCT Publications WO

94/12305, WO 94/11530, WO 97/29212 and WO 97/31256, the disclosures of which are incorporated herein by reference in their entireties.

Consequently, the invention concerns an array of nucleic acid molecules comprising at least one polynucleotide of the invention, particularly a probe or primer as described herein. Preferably, the invention concerns an array of nucleic acid comprising at least two polynucleotides of the invention, particularly probes or primers as described herein. Preferably, the invention concerns an array of nucleic acid comprising at least five polynucleotides of the invention, particularly probes or primers as described herein.

A preferred embodiment of the present invention is an array of polynucleotides of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 100, 500, 1000, 1500 or 2000 nucleotides in length which includes at least 1, 2, 5, 10, 15, 20, 35, 50, 100, 150 or 200 sequences selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, sequences fully complementary thereto, and fragments thereof.

Methods of making the polynucleotides of the invention

The present invention also comprises methods of making the polynucleotides of the invention, including the polynucleotides of SEQ ID Nos: 1-241, genomic DNA obtainable therefrom, or fragment thereof. These methods comprise sequentially linking together nucleotides to produce the nucleic acids having the preceding sequences. Polynucleotides of the invention may be synthesized either enzymatically using techniques well known to those skilled in the art including amplification or hybridization-based methods as described herein, or chemically.

A variety of chemical methods of synthesizing nucleic acids are known to those skilled in the art. In many of these methods, synthesis is conducted on a solid support. These included the 3' phosphoramidite methods in which the 3' terminal base of the desired oligonucleotide is immobilized on an insoluble carrier. The nucleotide base to be added is blocked at the 5' hydroxyl and activated at the 3' hydroxyl so as to cause coupling with the immobilized nucleotide base. Deblocking of the new immobilized nucleotide compound and repetition of the cycle will produce the desired polynucleotide. Alternatively, polynucleotides may be prepared as described in U.S. Patent No. 5,049,656, which disclosure is hereby incorporated by reference in its entirety. In some embodiments, several polynucleotides prepared as described above are ligated together to generate longer polynucleotides having a desired sequence.

POLYPEPTIDES OF THE INVENTION

The term "GENSET polypeptides" is used herein to embrace all of the proteins and polypeptides of the present invention. The present invention encompasses GENSET polypeptides, including recombinant, isolated or purified GENSET polypeptides consisting of, consisting essentially of, or comprising a sequence selected from the group consisting of SEQ ID Nos: 242-482, the polypeptides encoded by human cDNAs contained in the deposited clones, the mature

proteins included in SEQ ID Nos: 242-272 and 274-384, mature proteins encoded by the clone inserts of the deposited clone pool, and variants thereof. Other objects of the invention are polypeptides encoded by the polynucleotides of the invention as well as fusion polypeptides comprising such polypeptide.

Polypeptide variants

The present invention further provides for GENSET polypeptides encoded by allelic and splice variants, orthologs, and/or species homologues. Procedures known in the art can be used to obtain, allelic variants, splice variants, orthologs, and/or species homologues of polynucleotides encoding by polypeptides of the group consisting of SEQ ID Nos: 242-482, mature proteins included in SEQ ID Nos: 242-272 and 274-384, and polypeptides either fill-length or mature encoded by the clone inserts of the deposited clone pool, using information from the sequences disclosed herein or the clones deposited with the ATCC.

The polypeptides of the present invention also include polypeptides having an amino acid sequence at least 50% identical, more preferably at least 60% identical, and still more preferably 70%, 80%, 90%, 95%, 96%, 97%, 98% or 99% identical to a polypeptide selected from the group consisting of the sequences of SEQ ID Nos: 242-482, mature proteins included in sequences of SEQ ID Nos: 242-272 and 274-384, and full-length or mature polypeptides encoded by the clone inserts of the deposited clone pool. By a polypeptide having an amino acid sequence at least, for example, 95% "identical" to a query amino acid sequence of the present invention, it is intended that the amino acid sequence of the subject polypeptide is identical to the query sequence except that the subject polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the query amino acid sequence. In other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a query amino acid sequence, up to 5% (5 of 100) of the amino acid residues in the subject sequence may be inserted, deleted, (indels) or substituted with another amino acid.

Further polypeptides of the present invention include polypeptides which have at least 90% similarity, more preferably at least 95% similarity, and still more preferably at least 96%, 97%, 98% or 99% similarity to those described above. By a polypeptide having an amino acid sequence at least, for example, 95% "similar" to a query amino acid sequence of the present invention, it is intended that the amino acid sequence of the subject polypeptide is similar (i.e. contain identical or equivalent amino acid residues) to the query sequence except that the subject polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the query amino acid sequence. In other words, to obtain a polypeptide having an amino acid sequence at least 95% similar to a query amino acid sequence, up to 5% (5 of 100) of the amino acid residues in the subject sequence may be inserted, deleted, (indels) or substituted with another non-equivalent amino acid.

These alterations of the reference sequence may occur at the amino or carboxy terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence or in one or more contiguous groups within the reference sequence. The query sequence may be an entire amino acid sequence selected from the group consisting of sequences of SEQ ID Nos: 242-482 and those encoded by the clone inserts of the deposited clone pool or any fragment specified as described herein.

The variant polypeptides described herein are included in the present invention regardless of whether they have their normal biological activity. This is because even where a particular polypeptide molecule does not have biological activity, one of skill in the art would still know how to use the polypeptide, for instance, as a vaccine or to generate antibodies. Other uses of the polypeptides of the present invention that do not have GENSET biological activity include, *inter alia,* as epitope tags, in epitope mapping, and as molecular weight markers on SDS-PAGE gels or on molecular sieve gel filtration columns using methods known to those of skill in the art. As described below, the polypeptides of the present invention can also be used to raise polyclonal and monoclonal antibodies, which are useful in assays for detecting GENSET protein expression or as agonists and antagonists capable of enhancing or inhibiting GENSET protein function. Further, such polypeptides can be used in the yeast two-hybrid system to "capture" GENSET protein binding proteins, which are also candidate agonists and antagonists according to the present invention (*See, e.g.,* Fields *et al.* 1989), which disclosure is hereby incorporated by reference in its entirety.

Preparation of the polypeptides of the invention

The polypeptides of the present invention can be prepared in any suitable manner. Such polypeptides include isolated naturally occurring polypeptides, recombinantly produced polypeptides, synthetically produced polypeptides, or polypeptides produced by a combination of these methods. The polypeptides of the present invention are preferably provided in an isolated form, and may be partially or preferably substantially purified.

Consequently, the present invention also comprises methods of making the polypeptides of the invention, particularly polypeptides encoded by the cDNAs of SEQ ID Nos: 1-241, mature proteins encoded by fragments of SEQS ID Nos: 1-31 and 33-143, full-length and mature polypeptides encoded by the clone inserts of the deposited clone pool, genomic DNA obtainable therefrom, or fragments thereof and methods of making the polypeptides of SEQ ID Nos: 242-482, mature polypeptides included in SEQ ID Nos: 242-272 and 274-384, or fragments thereof. The methods comprise sequentially linking together amino acids to produce the nucleic polypeptides having the preceding sequences. In some embodiments, the polypeptides made by these methods are 150 amino acids or less in length. In other embodiments, the polypeptides made by these methods are 120 amino acids or less in length.

Isolation

From natural sources

The GENSET proteins of the invention may be isolated from natural sources, including bodily fluids, tissues and cells, whether directly isolated or cultured cells, of humans or non-human animals. Methods for extracting and purifying natural proteins are known in the art, and include the use of detergents or chaotropic agents to disrupt particles followed by differential extraction and separation of the polypeptides by ion exchange chromatography, affinity chromatography, sedimentation according to density, and gel electrophoresis. See, for example, "Methods in Enzymology, Academic Press, 1993" for a variety of methods for purifying proteins, which disclosure is hereby incorporated by reference in its entirety. Polypeptides of the invention also can be purified from natural sources using antibodies directed against the polypeptides of the invention, such as those described herein, in methods which are well known in the art of protein purification.

From recombinant sources

Preferably, the GENSET polypeptides of the invention are recombinantly produced using routine expression methods known in the art. The polynucleotide encoding the desired polypeptide is operably linked to a promoter into an expression vector suitable for any convenient host. Both eukaryotic and prokaryotic host systems are used in forming recombinant polypeptides. The polypeptide is then isolated from lysed cells or from the culture medium and purified to the extent needed for its intended use.

Any GENSET polynucleotide, including those described in SEQ ID Nos: 1-241, those of clone inserts of the deposited clone pool, and allelic variants thereof may be used to express GENSET polypeptides. The nucleic acid encoding the GENSET polypeptide to be expressed is operably linked to a promoter in an expression vector using conventional cloning technology. The GENSET insert in the expression vector may comprise the full coding sequence for the GENSET protein or a portion thereof, especially the sequence for a mture polypeptide. For example, the GENSET derived insert may encode a polypeptide comprising at least 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150 or 200 consecutive amino acids of a GENSET protein selected from the group consisting of sequences of SEQ ID Nos: 242-482 and polypeptides encoded by the clone inserts of the deposited clone pool.

Consequently, a further embodiment of the present invention is a method of making a polypeptide comprising a protein selected from the group consisting of sequences of SEQ ID Nos: 242-482 and polypeptides encoded by the clone inserts of the deposited clone pool, said method comprising the steps of

a) obtaining a cDNA comprising a sequence selected from the group consisting of i) the sequences SEQ ID Nos: 1-241, ii) the sequences of clone inserts of the deposited clone pool one, iii) sequences encoding one of the polypeptide of SEQ ID Nos: 242-482, and iv) sequences of

polynucleotides encoding a polypeptide which is encoded by one of the clone insert of the deposited clone pool;

b) inserting said cDNA in an expression vector such that the cDNA is operably linked to a promoter; and

c) introducing said expression vector into a host cell whereby said host cell produces said polypeptide.

In one aspect of this embodiment, the method further comprises the step of isolating the polypeptide. Another embodiment of the present invention is a polypeptide obtainable by the method described in the preceding paragraph.

The expression vector is any of the mammalian, yeast, insect or bacterial expression systems known in the art. Commercially available vectors and expression systems are available from a variety of suppliers including Genetics Institute (Cambridge, MA), Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence is optimized for the particular expression organism in which the expression vector is introduced, as explained in U.S. Patent No. 5,082,767, which disclosure is hereby incorporated by reference in its entirety.

In one embodiment, the entire coding sequence of a GENSET cDNA and the 3'UTR through the poly A signal of the cDNA is operably linked to a promoter in the expression vector. Alternatively, if the nucleic acid encoding a portion of the GENSET protein lacks a methionine to serve as the initiation site, an initiating methionine can be introduced next to the first codon of the nucleic acid using conventional techniques. Similarly, if the insert from the GENSET cDNA lacks a poly A signal, this sequence can be added to the construct by, for example, splicing out the Poly A signal from pSG5 (Stratagene) using BglI and SalI restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene). pXT1 contains the LTRs and a portion of the gag gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex Thymidine Kinase promoter and the selectable neomycin gene. The nucleic acid encoding the GENSET protein or a portion thereof is obtained by PCR from a vector containing a GENSET cDNA selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and the clone inserts of the deposited clone pool using oligonucleotide primers complementary to the GENSET cDNA or portion thereof and containing restriction endonuclease sequences for Pst I incorporated into the 5' primer and BglII at the 5' end of the corresponding cDNA 3' primer, taking care to ensure that the sequence encoding the GENSET protein or a portion thereof is positioned properly with respect to the poly A signal. The purified fragment obtained from the resulting PCR reaction is digested with PstI, blunt ended with an exonuclease, digested with Bgl II, purified and ligated to pXT1, now containing a poly A signal and digested with BglII.

Alternatively, cDNAs encoding secreted proteins may be cloned into pED6dpc2 (DiscoverEase, Genetics Institute, Cambridge, MA). The resulting pED6dpc2 constructs may be transfected into a suitable host cell, such as COS 1 cells. Methotrexate resistant cells are selected and expanded. Preferably, the secreted protein expressed from the cDNA is released into the culture medium thereby facilitating purification.

In another embodiment, it is often advantageous to add to the recombinant polynucleotide additional nucleotide sequence which codes for secretory or leader sequences, pro-sequences, sequences which aid in purification, such as multiple histidine residues, or an additional sequence for stability during recombinant production.

As a control, the expression vector lacking a cDNA insert is introduced into host cells or organisms.

Transfection of a GENSET expressing vector into mouse NTH 3T3 cells is but one embodiment of introducing polynucleotides into host cells. Introduction of a polynucleotide encoding a polypeptide into a host cell can be effected by calcium phosphate transfection, DEAE-dextran mediated transfection, cationic lipid-mediated transfection, electroporation, transduction, infection, or other methods. Such methods are described in many standard laboratory manuals, such as Davis *et al.* (1986), which disclosure is hereby incorporated by reference in its entirety. It is specifically contemplated that the polypeptides of the present invention may in fact be expressed by a host cell lacking a recombinant vector.

Recombinant cell extracts, or proteins from the culture medium if the expressed polypeptide is secreted, are then prepared and proteins separated by gel electrophoresis. If desired, the proteins may be ammonium sulfate precipitated or separated based on size or charge prior to electrophoresis. The proteins present are detected using techniques such as Coomassie or silver staining or using antibodies against the protein encoded by the GENSET cDNA of interest. Coomassie and silver staining techniques are familiar to those skilled in the art.

Proteins from the host cells or organisms containing an expression vector which contains the GENSET cDNA or a fragment thereof are compared to those from the control cells or organism. The presence of a band from the cells containing the expression vector which is absent in control cells indicates that the GENSET cDNA is expressed. Generally, the band corresponding to the protein encoded by the GENSET cDNA will have a mobility near that expected based on the number of amino acids in the open reading frame of the cDNA. However, the band may have a mobility different than that expected as a result of modifications such as glycosylation, ubiquitination, or enzymatic cleavage.

Alternatively, the GENSET polypeptide to be expressed may also be a product of transgenic animals, i.e., as a component of the milk of transgenic cows, goats, pigs or sheeps which are characterized by somatic or germ cells containing a nucleotide sequence encoding the protein of interest.

A polypeptide of this invention can be recovered and purified from recombinant cell cultures by well-known methods including differential extraction, ammonium sulfate or ethanol precipitation, acid extraction, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography and lectin chromatography. See, for example, "Methods in Enzymology", *supra* for a variety of methods for purifying proteins. Most preferably, high performance liquid chromatography ("HPLC") is employed for purification. A recombinantly produced version of a GENSET polypeptide can be substantially purified using techniques described herein or otherwise known in the art, such as, for example, by the one-step method described in Smith and Johnson (1988), which disclosure is hereby incorporated by reference in its entirety. Polypeptides of the invention also can be purified from recombinant sources using antibodies directed against the polypeptides of the invention, such as those described herein, in methods which are well known in the art of protein purification.

Preferably, the recombinantly expressed GENSET polypeptide is purified using standard immunochromatography techniques such as the one described in the section entitled "Immunoaffinity Chromatography". In such procedures, a solution containing the protein of interest, such as the culture medium or a cell extract, is applied to a column having antibodies against the protein attached to the chromatography matrix. The recombinant protein is allowed to bind the immunochromatography column. Thereafter, the column is washed to remove non-specifically bound proteins. The specifically bound protein is then released from the column and recovered using standard techniques.

If antibody production is not possible, the GENSET cDNA sequence or fragment thereof may be incorporated into expression vectors designed for use in purification schemes employing chimeric polypeptides. In such strategies the coding sequence of the GENSET cDNA or fragment thereof is inserted in frame with the gene encoding the other half of the chimera. The other half of the chimera may be beta-globin or a nickel binding polypeptide encoding sequence. A chromatography matrix having antibody to beta-globin or nickel attached thereto is then used to purify the chimeric protein. Protease cleavage sites may be engineered between the beta-globin gene or the nickel binding polypeptide and the GENSET cDNA or fragment thereof. Thus, the two polypeptides of the chimera may be separated from one another by protease digestion.

One useful expression vector for generating beta-globin chimerics is pSG5 (Stratagene), which encodes rabbit beta-globin. Intron II of the rabbit beta-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal incorporated into the construct increases the level of expression. These techniques as described are well known to those skilled in the art of molecular biology. Standard methods are published in methods texts such as Davis *et al.*, (1986) and many of the methods are available from Stratagene, Life Technologies, Inc., or Promega.

Polypeptide may additionally be produced from the construct using *in vitro* translation systems such as the *In vitro* Express™ Translation Kit (Stratagene).

Depending upon the host employed in a recombinant production procedure, the polypeptides of the present invention may be glycosylated or may be non-glycosylated. In addition, polypeptides of the invention may also include an initial modified methionine residue, in some cases as a result of host-mediated processes. Thus, it is well known in the art that the N-terminal methionine encoded by the translation initiation codon generally is removed with high efficiency from any protein after translation in all eukaryotic cells. While the N-terminal methionine on most proteins also is efficiently removed in most prokaryotes, for some proteins, this prokaryotic removal process is inefficient, depending on the nature of the amino acid to which the N-terminal methionine is covalently linked.

From chemical synthesis

In addition, polypeptides of the invention, especially short protein fragments, can be chemically synthesized using techniques known in the art (*See, e.g.*, Creighton, 1983; and Hunkapiller *et al.*, 1984), which disclosures are hereby incorporated by reference in their entireties. For example, a polypeptide corresponding to a fragment of a polypeptide sequence of the invention can be synthesized by use of a peptide synthesizer. A variety of methods of making polypeptides are known to those skilled in the art, including methods in which the carboxyl terminal amino acid is bound to polyvinyl benzene or another suitable resin. The amino acid to be added possesses blocking groups on its amino moiety and any side chain reactive groups so that only its carboxyl moiety can react. The carboxyl group is activated with carbodiimide or another activating agent and allowed to couple to the immobilized amino acid. After removal of the blocking group, the cycle is repeated to generate a polypeptide having the desired sequence. Alternatively, the methods described in U.S. Patent No. 5,049,656, which disclosure is hereby incorporated by reference in its entirety, may be used.

Furthermore, if desired, nonclassical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the polypeptide sequence. Non-classical amino acids include, but are not limited to, to the D-isomers of the common amino acids, 2,4-diaminobutyric acid, a-amino isobutyric acid, 4-aminobutyric acid, Abu, 2-amino butyric acid, g-Abu, e-Ahx, 6-amino hexanoic acid, Aib, 2-amino isobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxyproline, sarcosine, citrulline, homocitrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, b-alanine, fluoroamino acids, designer amino acids such as b-methyl amino acids, Ca-methyl amino acids, Na-methyl amino acids, and amino acid analogs in general. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

Modifications

The invention encompasses polypeptides which are differentially modified during or after translation, e.g., by glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, linkage to an antibody molecule or other cellular ligand, etc. Any of numerous chemical modifications may be carried out by known techniques, including but not limited, to specific chemical cleavage by cyanogen bromide, trypsin, chymotrypsin, papain, V8 protease, NaBH4; acetylation, formylation, oxidation, reduction; metabolic synthesis in the presence of tunicamycin; etc.

Additional post-translational modifications encompassed by the invention include, for example, e.g., N-linked or O-linked carbohydrate chains, processing of N-terminal or C-terminal ends), attachment of chemical moieties to the amino acid backbone, chemical modifications of N-linked or O-linked carbohydrate chains, and addition or deletion of an N-terminal methionine residue as a result of prokaryotic host cell expression. The polypeptides may also be modified with a detectable label, such as an enzymatic, fluorescent, isotopic or affinity label to allow for detection and isolation of the protein.

Also provided by the invention are chemically modified derivatives of the polypeptides of the invention which may provide additional advantages such as increased solubility, stability and circulating time of the polypeptide, or decreased immunogenicity. See U.S. Patent No: 4,179,337. The chemical moieties for derivatization may be selected See U.S. Patent NO: 4,179,337, which disclosure is hereby incorporated by reference in its entirety. The chemical moieties for derivatization may be selected from water soluble polymers such as polyethylene glycol, ethylene glycol/propylene glycol copolymers, carboxymethylcellulose, dextran, polyvinyl alcohol and the like. The polypeptides may be modified at random positions within the molecule, or at predetermined positions within the molecule and may include one, two, three or more attached chemical moieties.

The polymer may be of any molecular weight, and may be branched or unbranched. For polyethylene glycol, the preferred molecular weight is between about 1 kDa and about 100 kDa (the term "about" indicating that in preparations of polyethylene glycol, some molecules will weigh more, some less, than the stated molecular weight) for ease in handling and manufacturing. Other sizes may be used, depending on the desired therapeutic profile (e.g., the duration of sustained release desired, the effects, if any on biological activity, the ease in handling, the degree or lack of antigenicity and other known effects of the polyethylene glycol to a therapeutic protein or analog).

The polyethylene glycol molecules (or other chemical moieties) should be attached to the protein with consideration of effects on functional or antigenic domains of the protein. There are a number of attachment methods available to those skilled in the art, e.g., EP 0 401 384, (coupling PEG to G-CSF), and Malik *et al.* (1992) (reporting pegylation of GM-CSF using tresyl chloride), which disclosures are hereby incorporated by reference in their entireties. For example,

polyethylene glycol may be covalently bound through amino acid residues via a reactive group, such as, a free amino or carboxyl group. Reactive groups are those to which an activated polyethylene glycol molecule may be bound. The amino acid residues having a free amino group may include lysine residues and the N-terminal amino acid residues; those having a free carboxyl group may include aspartic acid residues glutamic acid residues and the C-terminal amino acid residue. Sulfhydryl groups may also be used as a reactive group for attaching the polyethylene glycol molecules. Preferred for therapeutic purposes is attachment at an amino group, such as attachment at the N-terminus or lysine group.

One may specifically desire proteins chemically modified at the N-terminus. Using polyethylene glycol as an illustration of the present composition, one may select from a variety of polyethylene glycol molecules (by molecular weight, branching, etc.), the proportion of polyethylene glycol molecules to protein (polypeptide) molecules in the reaction mix, the type of pegylation reaction to be performed, and the method of obtaining the selected N-terminally pegylated protein. The method of obtaining the N-terminally pegylated preparation (i.e., separating this moiety from other monopegylated moieties if necessary) may be by purification of the N-terminally pegylated material from a population of pegylated protein molecules. Selective proteins chemically modified at the N-terminus modification may be accomplished by reductive alkylation, which exploits differential reactivity of different types of primary amino groups (lysine versus the N-terminal) available for derivatization in a particular protein. Under the appropriate reaction conditions, substantially selective derivatization of the protein at the N-terminus with a carbonyl group containing polymer is achieved.

Multimerization

The polypeptides of the invention may be in monomers or multimers (i.e., dimers, trimers, tetramers and higher multimers). Accordingly, the present invention relates to monomers and multimers of the polypeptides of the invention, their preparation, and compositions containing them. In specific embodiments, the polypeptides of the invention are monomers, dimers, trimers or tetramers. In additional embodiments, the multimers of the invention are at least dimers, at least trimers, or at least tetramers.

Multimers encompassed by the invention may be homomers or heteromers. As used herein, the term "homomer", refers to a multimer containing only polypeptides corresponding to the amino acid sequences of SEQ ID Nos: 242-482 or encoded by the clone inserts of the deposited clone pool (including fragments, variants, splice variants, and fusion proteins, corresponding to these polypeptides as described herein). These homomers may contain polypeptides having identical or different amino acid sequences. In a specific embodiment, a homomer of the invention is a multimer containing only polypeptides having an identical amino acid sequence. In another specific embodiment, a homomer of the invention is a multimer containing polypeptides having different

amino acid sequences. In specific embodiments, the multimer of the invention is a homodimer (*e.g.*, containing polypeptides having identical or different amino acid sequences) or a homotrimer (*e.g.*, containing polypeptides having identical and/or different amino acid sequences). In additional embodiments, the homomenc multimer of the invention is at least a homodimer, at least a homotrimer, or at least a homotetramer.

As used herein, the term "heteromer" refers to a multimer containing one or more heterologous polypeptides *(i.e.*, polypeptides of different proteins) in addition to the polypeptides of the invention. In a specific embodiment, the multimer of the invention is a heterodimer, a heterotrimer, or a heterotetramer. In additional embodiments, the heteromeric multimer of the invention is at least a heterodimer, at least a heterotrimer, or at least a heterotetramer.

Multimers of the invention may be the result of hydrophobic, hydrophilic, ionic and/or covalent associations and/or may be indirectly linked, by for example, liposome formation. Thus, in one embodiment, multimers of the invention, such as, for example, homodimers or homotrimers, are formed when polypeptides of the invention contact one another in solution. In another embodiment, heteromultimers of the invention, such as, for example, heterotrimers or heterotetramers, are formed when polypeptides of the invention contact antibodies to the polypeptides of the invention (including antibodies to the heterologous polypeptide sequence in a fusion protein of the invention) in solution. In other embodiments, multimers of the invention are formed by covalent associations with and/or between the polypeptides of the invention. Such covalent associations may involve one or more amino acid residues contained in the polypeptide sequence (e.g., that recited in the sequence listing, or contained in the polypeptide encoded by a deposited clone). In one instance, the covalent associations are cross-linking between cysteine residues located within the polypeptide sequences, which interact in the native (i.e., naturally occurring) polypeptide. In another instance, the covalent associations are the consequence of chemical or recombinant manipulation. Alternatively, such covalent associations may involve one or more amino acid residues contained in the heterologous polypeptide sequence in a fusion protein of the invention.

In one example, covalent associations are between the heterologous sequence contained in a fusion protein of the invention (see, e.g., US Patent Number 5,478,925, which disclosure is hereby incorporated by reference in its entirety). In a specific example, the covalent associations are between the heterologous sequence contained in an Fc fusion protein of the invention (as described herein). In another specific example, covalent associations of fusion proteins of the invention are between heterologous polypeptide sequence from another protein that is capable of forming covalently associated multimers, such as for example, oseteoprotegerin (see, e.g., International Publication No: WO 98/49305, the contents of which are herein incorporated by reference in its entirety). In another embodiment, two or more polypeptides of the invention are joined through peptide linkers. Examples include those peptide linkers described in U.S. Pat. No. 5,073,627

(hereby incorporated by reference). Proteins comprising multiple polypeptides of the invention separated by peptide linkers may be produced using conventional recombinant DNA technology.

Another method for preparing multimer polypeptides of the invention involves use of polypeptides of the invention fused to a leucine zipper or isoleucine zipper polypeptide sequence. Leucine zipper and isoleucine zipper domains are polypeptides that promote multimerization of the proteins in which they are found. Leucine zippers were originally identified in several DNA-binding proteins, and have since been found in a variety of different proteins (Landschulz *et al.*, 1988). Among the known leucine zippers are naturally occurring peptides and derivatives thereof that dimerize or trimerize. Examples of leucine zipper domains suitable for producing soluble multimeric proteins of the invention are those described in PCT application WO 94/10308, hereby incorporated by reference. Recombinant fusion proteins comprising a polypeptide of the invention fused to a polypeptide sequence that dimerizes or trimerizes in solution are expressed in suitable host cells, and the resulting soluble multimeric fusion protein is recovered from the culture supernatant using techniques known in the art.

Trimeric polypeptides of the invention may offer the advantage of enhanced biological activity. Preferred leucine zipper moieties and isoleucine moieties are those that preferentially form trimers. One example is a leucine zipper derived from lung surfactant protein D (SPD), as described in Hoppe *et al.* (1994) and in U.S. patent application Ser. No. 08/446,922, which disclosure is hereby incorporated by reference in its entirety. Other peptides derived from naturally occurring trimeric proteins may be employed in preparing trimeric polypeptides of the invention. In another example, proteins of the invention are associated by interactions between Flag® polypeptide sequence contained in fusion proteins of the invention containing Flag® polypeptide sequence. In a further embodiment, associations proteins of the invention are associated by interactions between heterologous polypeptide sequence contained in Flag® fusion proteins of the invention and anti Flag® antibody.

The multimers of the invention may be generated using chemical techniques known in the art. For example, polypeptides desired to be contained in the multimers of the invention may be chemically cross-linked using linker molecules and linker molecule length optimization techniques known in the art (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety). Additionally, multimers of the invention may be generated using techniques known in the art to form one or more inter-molecule cross-links between the cysteine residues located within the sequence of the polypeptides desired to be contained in the multimer (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety). Further, polypeptides of the invention may be routinely modified by the addition of cysteine or biotin to the C terminus or N-terminus of the polypeptide and techniques known in the art may be applied to generate multimers containing one or more of these modified polypeptides (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety). Additionally,

30 techniques known in the art may be applied to generate liposomes containing the polypeptide components desired to be contained in the multimer of the invention (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety).

Alternatively, multimers of the invention may be generated using genetic engineering techniques known in the art. In one embodiment, polypeptides contained in multimers of the invention are produced recombinantly using fusion protein technology described herein or otherwise known in the art (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety). In a specific embodiment, polynucleotides coding for a homodimer of the invention are generated by ligating a polynucleotide sequence encoding a polypeptide of the invention to a sequence encoding a linker polypeptide and then further to a synthetic polynucleotide encoding the translated product of the polypeptide in the reverse orientation from the original C-terminus to the N-terminus (lacking the leader sequence) (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety). In another embodiment, recombinant techniques described herein or otherwise known in the art are applied to generate recombinant polypeptides of the invention which contain a transmembrane domain (or hydrophobic or signal peptide) and which can be incorporated by membrane reconstitution techniques into liposomes (see, e.g., US Patent Number 5,478,925, which is herein incorporated by reference in its entirety).

Mutated polypeptides

To improve or alter the characteristics of GENSET polypeptides of the present invention, protein engineering may be employed. Recombinant DNA technology known to those skilled in the art can be used to create novel mutant proteins or muteins including single or multiple amino acid substitutions, deletions, additions, or fusion proteins. Such modified polypeptides can show, e.g., increased/decreased biological activity or increased/decreased stability. In addition, they may be purified in higher yields and show better solubility than the corresponding natural polypeptide, at least under certain purification and storage conditions. Further, the polypeptides of the present invention may be produced as multimers including dimers, trimers and tetramers. Multimerization may be facilitated by linkers or recombinantly though heterologous polypeptides such as Fc regions.

N- and C-terminal deletions

It is known in the art that one or more amino acids may be deleted from the N-terminus or C-terminus without substantial loss of biological function. For instance, Ron *et al.* (1993), reported modified KGF proteins that had heparin binding activity even if 3, 8, or 27 N-terminal amino acid residues were missing. Accordingly, the present invention provides polypeptides having one or more residues deleted from the amino terminus of the polypeptides of SEQ ID Nos: 242-482 or that encoded by the clone inserts of the deposited clone pool. Similarly, many examples of biologically functional C-terminal deletion mutants are known. For instance, Interferon gamma shows up to ten times higher activities by deleting 810 amino acid residues from the C-terminus of the protein (*See,*

e.g., Dobeli, *et al.* 1988), which disclosure is hereby incorporated by reference in its entirety. Accordingly, the present invention provides polypeptides having one or more residues deleted from the carboxy terminus of the polypeptides shown of SEQ ID Nos: 242-482 or encoded by the clone inserts of the deposited clone pool. The invention also provides polypeptides having one or more amino acids deleted from both the amino and the carboxyl termini as described below.

Other mutations

Other mutants in addition to N- and C-terminal deletion forms of the protein discussed above are included in the present invention. It also will be recognized by one of ordinary skill in the art that some amino acid sequences of the GENSET polypeptides of the present invention can be varied without significant effect of the structure or function of the protein. If such differences in sequence are contemplated, it should be remembered that there will be critical areas on the protein which determine activity. Thus, the invention further includes variations of the GENSET polypeptides which show substantial GENSET polypeptide activity. Such mutants include deletions, insertions, inversions, repeats, and substitutions selected according to general rules known in the art so as to have little effect on activity. For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided.

There are two main approaches for studying the tolerance of an amino acid sequence to change (See, Bowie *et al.* 1994), which disclosure is hereby incorporated by reference in its entirety. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection.

The second approach uses genetic engineering to introduce amino acid changes at specific positions of a cloned gene and selections or screens to identify sequences that maintain functionality. These studies have revealed that proteins are surprisingly tolerant of amino acid substitutions. The studies indicate which amino acid changes are likely to be permissive at a certain position of the protein. For example, most buried amino acid residues require nonpolar side chains, whereas few features of surface side chains are generally conserved. Other such phenotypically silent substitutions are described by Bowie *et al.* (supra) and the references cited therein.

Typically seen as conservative substitutions are the replacements, one for another, among the aliphatic amino acids Ala, Val, Leu and Phe; interchange of the hydroxyl residues Ser and Thr, exchange of the acidic residues Asp and Glu, substitution between the amide residues Asn and Gln, exchange of the basic residues Lys and Arg and replacements among the aromatic residues Phe, Tyr. Thus, the fragment, derivative, analog, or homologue of the polypeptide of the present invention may be, for example: (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code: or (ii) one in which one or more of the amino acid residues includes a substituent group: or

(iii) one in which the GENSET polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol): or (iv) one in which the additional amino acids are fused to the above form of the polypeptide, such as an IgG Fc fusion region peptide or leader or secretory sequence or a sequence which is employed for purification of the above form of the polypeptide or a pro-protein sequence. Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

Thus, the GENSET polypeptides of the present invention may include one or more amino acid substitutions, deletions, or additions, either from natural mutations or human manipulation. As indicated, changes are preferably of a minor nature, such as conservative amino acid substitutions that do not significantly affect the folding or activity of the protein. The following groups of amino acids generally represent equivalent changes: (1) Ala, Pro, Gly, Glu, Asp, Gln, Asn, Ser, Thr; (2) Cys, Ser, Tyr, Thr; (3) Val, Ile, Leu, Met, Ala, Phe; (4) Lys, Arg, His; (5) Phe, Tyr, Trp, His.

A specific embodiment of a modified GENSET peptide molecule of interest according to the present invention, includes, but is not limited to, a peptide molecule which is resistant to proteolysis, is a peptide in which the -CONH- peptide bond is modified and replaced by a (CH2NH) reduced bond, a (NHCO) retro inverso bond, a (CH2-O) methylene-oxy bond, a (CH2-S) thiomethylene bond, a (CH2CH2) carba bond, a (CO-CH2) cetomethylene bond, a (CHOH-CH2) hydroxyethylene bond), a (N-N) bound, a E-alcene bond or also a -CH=CH- bond. The invention also encompasses a human GENSET polypeptide or a fragment or a variant thereof in which at least one peptide bond has been modified as described above.

Amino acids in the GENSET proteins of the present invention that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (*See, e.g.*, Cunningham *et al.* 1989), which disclosure is hereby incorporated by reference in its entirety. The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity using assays appropriate for measuring the function of the particular protein. Of special interest are substitutions of charged amino acids with other charged or neutral amino acids which may produce proteins with highly desirable improved characteristics, such as less aggregation. Aggregation may not only reduce activity but also be problematic when preparing pharmaceutical formulations, because aggregates can be immunogenic, (*See, e.g.*, Pinckard *et al.*, 1967; Robbins, *et al.*, 1987; and Cleland, *et al.*, 1993).

A further embodiment of the invention relates to a polypeptide which comprises the amino acid sequence of a GENSET polypeptide having an amino acid sequence which contains at least one conservative amino acid substitution, but not more than 50 conservative amino acid substitutions, not more than 40 conservative amino acid substitutions, not more than 30 conservative amino acid substitutions, and not more than 20 conservative amino acid substitutions. Also provided are

polypeptides which comprise the amino acid sequence of a GENSET polypeptide, having at least one, but not more than 10, 9, 8, 7, 6, 5, 4, 3, 2 or 1 conservative amino acid substitutions.

Polypeptide fragments

Structural definition

The present invention is further directed to fragments of the amino acid sequences described herein such as the polypeptides of SEQ ID Nos: 242-482, mature polypeptides included in SEQ ID Nos: 242-272 and 274-384, or full-length or mature polypeptides encoded by the clone inserts of the deposited clone pool. More specifically, the present invention embodies purified, isolated, and recombinant polypeptides comprising at least 6, preferably at least 8 to 10, more preferably 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450 or 500 consecutive amino acids of a polypeptide selected from the group consisting of the sequences of SEQ ID Nos: 242-482, mature polypeptides included in SEQ ID Nos: 242-272 and 274-384, and full-length or mature polypeptides encoded by the clone inserts of the deposited clone pool, and other polypeptides of the present invention.

In addition to the above polypeptide fragments, further preferred sub-genuses of polypeptides comprise at least 6 amino acids, wherein "at least 6" is defined as any integer between 6 and the integer representing the C-terminal amino acid of the polypeptide of the present invention including the polypeptide sequences of the sequence listing below. Further included are species of polypeptide fragments at least 6 amino acids in length, as described above, that are further specified in terms of their N-terminal and C-terminal positions. However, included in the present invention as individual species are all polypeptide fragments, at least 6 amino acids in length, as described above, and may be particularly specified by a N-terminal and C-terminal position. That is, every combination of a N-terminal and C-terminal position that a fragment at least 6 contiguous amino acid residues in length could occupy, on any given amino acid sequence of the sequence listing or of the present invention is included in the present invention

The present invention also provides for the exclusion of any fragment species specified by N-terminal and C-terminal positions or of any fragment sub-genus specified by size in amino acid residues as described above. Any number of fragments specified by N-terminal and C-terminal positions or by size in amino acid residues as described above may be excluded as individual species.

The above polypeptide fragments of the present invention can be immediately envisaged using the above description and are therefore not individually listed solely for the purpose of not unnecessarily lengthening the specification. Moreover, the above fragments need not have a GENSET biological activity, although polypeptides having these activities are preferred embodiments of the invention, since they would be useful, for example, in immunoassays, in epitope mapping, epitope tagging, as vaccines, and as molecular weight markers. The above

fragments may also be used to generate antibodies to a particular portion of the polypeptide. These antibodies can then be used in immunoassays well known in the art to distinguish between human and non-human cells and tissues or to determine whether cells or tissues in a biological sample are or are not of the same type which express the polypeptides of the present invention.

It is noted that the above species of polypeptide fragments of the present invention may alternatively be described by the formula "a to b"; where "a" equals the N-terminal most amino acid position and "b" equals the C-terminal most amino acid position of the polynucleotide; and further where "a" equals an integer between 1 and the number of amino acids of the polypeptide sequence of the present invention minus 6, and where "b" equals an integer between 7 and the number of amino acids of the polypeptide sequence of the present invention; and where "a" is an integer smaller then "b" by at least 6.

The present invention also provides for the exclusion of any species of polypeptide fragments of the present invention specified by 5' and 3' positions or sub-genuses of polypeptides specified by size in amino acids as described above. Any number of fragments specified by 5' and 3' positions or by size in amino acids, as described above, may be excluded. Specifically excluded from the invention are the polypeptide fragments encoded by the preferentially excluded polynucleotide fragments described in Table IV, and in Tables Va and Vb. Table IV and Tables Va and Vb provide for the exclusion of polypeptides, independently from each other, in addition to those described elsewhere in the specification and is therefore, not meant as limiting description.

Functional definition

Preferred polypeptide fragments of the invention are isolated, purified or recombinant polypeptides comprising, consisting of, or consisting essentially of signal peptides, preferably signal peptides selected from the group consisting of SEQ ID Nos: 242-272 and 274-384, signal peptides encoded by sequences of SEQ ID Nos: 1-31 and 33-143 and those encoded by the clone inserts of the deposited clone pool. Such polypeptides fragments are useful to design secretion vectors as described elsewhere in the application.

Other preferred polypeptide fragments of the invention are isolated, purified or recombinant polypeptides comprising, consisting of, or consisting essentially of mature proteins, preferably mature proteins selected from the group consisting of SEQ ID Nos: 242-272 and 274-384, mature proteins encoded by sequences of SEQ ID Nos: 1-31 and 33-143 and those encoded by the clone inserts of the deposited clone pool.

Domains

Preferred polynucleotide fragments of the invention are domains of polypeptides of the invention. Such domains may eventually comprise linear or structural motifs and signatures including, but not limited to, leucine zippers, helix-turn-helix motifs, post-translational modification sites such as glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences

encoding signal peptides which direct the secretion of the encoded proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites. Such domains may present a particular biological activity such as DNA or RNA-binding, secretion of proteins, transcription regulation, enzymatic activity, substrate binding activity, etc...

A domain has a size generally comprised between 3 and 2000 amino acids. In preferred embodiment, domains comprise a number of amino acids that is any integer between 6 and 500. Domains may be synthesized using any methods known to those skilled in the art, including those disclosed herein, particularly in the section entitled "Preparation of the polypeptides of the invention". Methods for determining the amino acids which make up a domain with a particular biological activity include mutagenesis studies and assays to determine the biological activity to be tested.

Alternatively, the polypeptides of the invention may be scanned for motifs, domains and/or signatures in databases using any computer method known to those skilled in the art. Searchable databases include Prosite (Hofmann *et al.*, 1999; Bucher and Bairoch 1994), Pfam (Sonnhammer *et al.*, 1997; Henikoff *et al.*, 2000; Bateman *et al.*, 2000), Blocks (Henikoff *et al.*, 2000), Print (Attwood *et al.*, 1996), Prodom (Sonnhammer and Kahn, 1994; Corpet *et al.* 2000), Sbase (Pongor *et al.*, 1993; Murvai *et al.*, 2000), Smart (Schultz *et al.*, 1998), Dali/FSSP (Holm and Sander, 1996, 1997 and 1999), HSSP (Sander and Schneider 1991), CATH (Orengo *et al.*, 1997; Pearl *et al.*, 2000), SCOP (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000), COG (Tatusov *et al.*, 1997 and 2000), specific family databases and derivatives thereof (Nevill-Manning *et al.*, 1998; Yona *et al.*, 1999; Attwood *et al.*, 2000), each of which disclosures are hereby incorporated by reference in their entireties. For a review on available databases, see issue 1 of volume 28 of Nucleic Acid Research (2000), which disclosure is hereby incorporated by reference in its entirety.

The polypeptides of SEQ ID NOs : 242-482 were screened for the presence of known structural or functional motifs or for the presence of signatures, small amino acid sequences that are well conserved amongst the members of a protein family. The search was conducted on the Pfam 5.5 database using HMMER-2.1.1 (for info see Sonnhammer et Durbin, http:/www.sanger.ac.uk/Pfam/), on a Blocks Plus database containing Blocks version 12.0, Prints version 26.0, Pfam version 5.3, Prodom version 99.1, and Domo version 2.0 using emotif (for info see Nevill-Manning *et al.*, *PNAS*, 95, 5865-5871, (1998), http://motif.stanford/edu/EMOTIF) and on the Prosite 16.0 database using bla (Tatusov, R. L. & Koonin, E. V. CABIOS 10, No. 4) and pfscan (http://www.isrec.isb-sib.ch/cgi-bin/man.cgi?section=1&topic=pfscan). Some of these predicted domains are described in Table VI. For these polypeptides referred to by their sequence identification numbers (column entitled "Seq Id No"), Table VI gives the designation of the domain (column entitled "Designation of domain") according to the database of domains indicated in the column entitled "Database " and the positions of preferred fragments within these sequences (column

entitled "Positions of domains"). Each fragment is represented by a-b where a and b are the start and end positions respectively of a given preferred fragment on the full-length polypeptide. Preferred fragments are separated from each other by a coma. As used herein, the term "domain described in Table VI" refers to all the domains listed in Table VI for a given GENSET protein referred to by its sequence identification number in the first column. It should be noted that in Table VI, the first methionine encountered is designated as amino acid number 1, i.e;, the leader sequence is not numbered negatively. In the appended sequence listing, the first amino acid of the mature protein resulting from cleavage of the signal peptide is designated as amino acid number 1 and the first amino acid of the signal peptide is designated with the appropriate negative number, in accordance with the regulations governing sequence listings.

Consequently, preferred polynucleotide fragments of the invention are domains of the polypeptides of SEQ ID Nos: 242-482. Therefore, the present invention encompasses isolated, purified, or recombinant polypeptides which consist of, consist essentially of, or comprise a contiguous span of at least 6, preferably at least 8 to 10, more preferably 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450 or 500 amino acids of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, where said contiguous span comprises at least 1, 2, 3, 5, or 10 amino acids positions of a domain described in Table VI of said selected sequence. The present invention also encompasses isolated, purified, or recombinant polypeptides comprising, consisting essentially of, or consisting of a contiguous span of at least 6, preferably at least 8 to 10, more preferably 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450 or 500 amino acids of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, where said contiguous span is a domain described in Table VI of said selected sequence. The present invention also encompasses isolated, purified, or recombinant polypeptides which comprise, consist of or consist essentially of a domain described in Table VI of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482 .

Polypeptides of the present invention that are not specifically described in this table are not considered as not belonging to a domain. This is because they may still be not recognized as such by the particular algorithms used or not be included in the particular database searched. In fact, all fragments of the polypeptides of the present invention, at least 6 amino acids residues in length, are included in the present invention as being a domain. Amino acid residues comprising other domains may be determined by looking in other databases than the ones currently cited to establish Table VI. The domains of the present invention preferably comprises 6 to 200 amino acids (i.e. any integer between 6 and 200, inclusive) of a polypeptide of the present invention. Also, included in the present invention are domain fragments between the integers of 6 and the full length GENSET

sequence of the sequence listing. All combinations of sequences between the integers of 6 and the full-length sequence of a GENSET polypeptide are included. The domain fragments may be specified by either the number of contiguous amino acid residues (as a sub-genus) or by specific N-terminal and C-terminal positions (as species) as described above for the polypeptide fragments of the present invention. Any number of domain fragments of the present invention may also be excluded in the same manner.

Epitopes and Antibody Fusions:

A preferred embodiment of the present invention is directed to epitope-bearing polypeptides and epitope-bearing polypeptide fragments. These epitopes may be "antigenic epitopes" or both an "antigenic epitope" and an "immunogenic epitope". An "immunogenic epitope" is defined as a part of a protein that elicits an antibody response *in vivo* when the polypeptide is the immunogen. On the other hand, a region of polypeptide to which an antibody binds is defined as an "antigenic determinant" or "antigenic epitope." The number of immunogenic epitopes of a protein generally is less than the number of antigenic epitopes (*See, e.g.,* Geysen, *et al.*, 1984), which disclosure is hereby incorporated by reference in its entirety. It is particularly noted that although a particular epitope may not be immunogenic, it is nonetheless useful since antibodies can be made to both immunogenic and antigenic epitopes.

An epitope can comprise as few as 3 amino acids in a spatial conformation, which is unique to the epitope. Generally an epitope consists of at least 6 such amino acids, and more often at least 8-10 such amino acids. In preferred embodiment, antigenic epitopes comprise a number of amino acids that is any integer between 3 and 50. Fragments which function as epitopes may be produced by any conventional means (*See, e.g.,* Houghten, 1985), also further described in U.S. Patent No. 4,631,21, which disclosures are hereby incorporated by reference in their entireties. Methods for determining the amino acids which make up an epitope include x-ray crystallography, 2-dimensional nuclear magnetic resonance, and epitope mapping, e.g., the Pepscan method described by Geysen *et al.* (1984); PCT Publication No. WO 84/03564; and PCT Publication No. WO 84/03506, which disclosures are hereby incorporated by reference in their entireties. Another example is the algorithm of Jameson and Wolf, (1988) (said reference incorporated by reference in its entirety). The Jameson-Wolf antigenic analysis, for example, may be performed using the computer program PROTEAN, using default parameters (Version 4.0 Windows, DNASTAR, Inc., 1228 South Park Street Madison, WI.

Antigenic epitopes predicted by the Jameson-Wolf algorithm for the polypeptides of SEQ ID Nos: 242-482 are presented in Table VII. For each GENSET polypeptide referred to by its sequence identification number in the column entitled "Seq Id No", a list of antigenic epitopes is given in the column entitled "Epitopes", each epitope being separated by a coma. Each fragment is represented by a-b where a and b are the start and end positions respectively of a given preferred

fragment. It should be noted that in Table VII, the first methionine encountered is designated as amino acid number 1, i.e; the leader sequence is not numbered negatively. In the appended sequence listing, the first amino acid of the mature protein resulting from cleavage of the signal peptide is designated as amino acid number 1 and the first amino acid of the signal peptide is designated with the appropriate negative number, in accordance with the regulations governing sequence listings. As used herein, the term "epitope described in Table VII" refers to all preferred polynucleotide fragments described in the second column of Table VII for a GENSET polypeptide referred to by its sequence identification number in the first column. It is pointed out that the immunogenic epitopes listed in Table VII describe only amino acid residues comprising epitopes predicted to have the highest degree of immunogenicity by a particular algorithm. Polypeptides of the present invention that are not specifically described as immunogenic are not considered non-antigenic. This is because they may still be antigenic *in vivo* but merely not recognized as such by the particular algorithm used. Alternatively, the polypeptides are most likely antigenic *in vitro* using methods such a phage display. Thus, listed in Table VII are the amino acid residues comprising only preferred epitopes, not a complete list. In fact, all fragments of the polypeptides of the present invention, at least 6 amino acids residues in length, are included in the present invention as being useful as antigenic epitope. Amino acid residues comprising other immunogenic epitopes may be determined by algorithms similar to the Jameson-Wolf analysis or by *in vivo* testing for an antigenic response using the methods described herein or those known in the art.

Therefore, the present invention encompasses isolated, purified, or recombinant polypeptides which consist of, consist essentially of, or comprise a contiguous span of at least 6, preferably at least 8 to 10, more preferably 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450 or 500 amino acids of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, where said contiguous span comprises at least 1, 2, 3, 5, or 10 amino acids positions of an epitope described in Table VII of said selected sequence. The present invention also encompasses isolated, purified, or recombinant polypeptides comprising, consisting essentially of, or consisting of a contiguous span of at least 6, preferably at least 8 to 10, more preferably 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450 or 500 amino acids of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482, to the extent that a contiguous span of these lengths is consistent with the lengths of said selected sequence, where said contiguous span is an epitope described in Table VII of said selected sequence. The present invention also encompasses isolated, purified, or recombinant polypeptides which comprise, consist of or consist essentially of an epitope described in Table VII of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482.

The epitope-bearing fragments of the present invention preferably comprises 6 to 50 amino acids (i.e. any integer between 6 and 50, inclusive) of a polypeptide of the present invention. Also, included in the present invention are antigenic fragments between the integers of 6 and the full length GENSET sequence of the sequence listing. All combinations of sequences between the integers of 6 and the full-length sequence of a GENSET polypeptide are included. The epitope-bearing fragments may be specified by either the number of contiguous amino acid residues (as a sub-genus) or by specific N-terminal and C-terminal positions (as species) as described above for the polypeptide fragments of the present invention. Any number of epitope-bearing fragments of the present invention may also be excluded in the same manner.

Antigenic epitopes are useful, for example, to raise antibodies, including monoclonal antibodies that specifically bind the epitope (See, Wilson *et al.*, 1984; and Sutcliffe, *et al.*, 1983), which disclosures are hereby incorporated by reference in their entireties. The antibodies are then used in various techniques such as diagnostic and tissue/cell identification techniques, as described herein, and in purification methods such as immunoaffinity chromatography.

An antibody or other compound that specifically binds to a polypeptide or polynucleotide of the invention is also said to "selectively recognize" the polypeptide or polynucleotide.

Similarly, immunogenic epitopes can be used to induce antibodies according to methods well known in the art (See, Sutcliffe *et al.*, supra; Wilson *et al.*, supra; Chow *et al.*;(1985) and Bittle, *et al.*, (1985), which disclosures are hereby incorporated by reference in their entireties). A preferred immunogenic epitope includes the natural GENSET protein. The immunogenic epitopes may be presented together with a carrier protein, such as an albumin, to an animal system (such as rabbit or mouse) or, if it is long enough (at least about 25 amino acids), without a carrier. However, immunogenic epitopes comprising as few as 8 to 10 amino acids have been shown to be sufficient to raise antibodies capable of binding to, at the very least, linear epitopes in a denatured polypeptide (e.g., in Western blotting.).

Epitope-bearing polypeptides of the present invention are used to induce antibodies according to methods well known in the art including, but not limited to, *in vivo* immunization, *in vitro* immunization, and phage display methods (*See, e.g.,* Sutcliffe, *et al.*, *supra*; Wilson, *et al.*, *supra*, and Bittle, *et al.*, *supra*). If *in vivo* immunization is used, animals may be immunized with free peptide; however, anti-peptide antibody titer may be boosted by coupling of the peptide to a macromolecular carrier, such as keyhole limpet hemacyanin (KLH) or tetanus toxoid. For instance, peptides containing cysteine residues may be coupled to a carrier using a linker such as -maleimidobenzoyl- N-hydroxysuccinimide ester (MBS), while other peptides may be coupled to carriers using a more general linking agent such as glutaraldehyde. Animals such as rabbits, rats and mice are immunized with either free or carrier-coupled peptides, for instance, by intraperitoneal and/or intradermal injection of emulsions containing about 100 µgs of peptide or carrier protein and Freund's adjuvant. Several booster injections may be needed, for instance, at intervals of about two

weeks, to provide a useful titer of anti-peptide antibody, which can be detected, for example, by ELISA assay using free peptide adsorbed to a solid surface. The titer of anti-peptide antibodies in serum from an immunized animal may be increased by selection of anti-peptide antibodies, for instance, by adsorption to the peptide on a solid support and elution of the selected antibodies according to methods well known in the art.

As one of skill in the art will appreciate, and discussed above, the polypeptides of the present invention comprising an immunogenic or antigenic epitope can be fused to heterologous polypeptide sequences. For example, the polypeptides of the present invention may be fused with the constant domain of immunoglobulins (IgA, IgE, IgG, IgM), or portions thereof (CH1, CH2, CH3, any combination thereof including both entire domains and portions thereof) resulting in chimeric polypeptides. These fusion proteins facilitate purification, and show an increased half-life *in vivo*. This has been shown, *e.g.*, for chimeric proteins consisting of the first two domains of the human CD4-polypeptide and various domains of the constant regions of the heavy or light chains of mammalian immunoglobulins (*See, e.g.*, EPA 0,394,827; and Traunecker *et al.*, 1988), which disclosures are hereby incorporated by reference in their entireties. Fusion proteins that have a disulfide-linked dimeric structure due to the IgG portion can also be more efficient in binding and neutralizing other molecules than monomeric polypeptides or fragments thereof alone (*See, e.g.*, Fountoulakis *et al.*, 1995), which disclosure is hereby incorporated by reference in its entirety. Nucleic acids encoding the above epitopes can also be recombined with a gene of interest as an epitope tag to aid in detection and purification of the expressed polypeptide.

Additional fusion proteins of the invention may be generated through the techniques of gene-shuffling, motif-shuffling, exon-shuffling, or codon-shuffling (collectively referred to as "DNA shuffling"). DNA shuffling may be employed to modulate the activities of polypeptides of the present invention thereby effectively generating agonists and antagonists of the polypeptides. See, for example, U.S. Patent Nos.: 5,605,793; 5,811,238; 5,834,252; 5,837,458; and Patten, *et al.*, (1997); Harayama, (1998); Hansson, *et al* (1999); and Lorenzo and Blasco, (1998). (Each of these documents are hereby incorporated by reference). In one embodiment, one or more components, motifs, sections, parts, domains, fragments, etc., of coding polynucleotides of the invention, or the polypeptides encoded thereby may be recombined with one or more components, motifs, sections, parts, domains, fragments, etc. of one or more heterologous molecules.

The present invention further encompasses any combination of the polypeptide fragments listed in this section.

Antibodies:

Definitions

The present invention further relates to antibodies and T-cell antigen receptors (TCR), which specifically bind the polypeptides, and more specifically, the epitopes of the polypeptides of

the present invention. The antibodies of the present invention include IgG (including IgG1, IgG2, IgG3, and IgG4), IgA (including IgA1 and IgA2), IgD, IgE, or IgM, and IgY. The term "antibody" (Ab) refers to a polypeptide or group of polypeptides which are comprised of at least one binding domain, where a binding domain is formed from the folding of variable domains of an antibody molecule to form three-dimensional binding spaces with an internal surface shape and charge distribution complementary to the features of an antigenic determinant of an antigen, which allows an immunological reaction with the antigen. As used herein, the term "antibody" is meant to include whole antibodies, including single-chain whole antibodies, and antigen binding fragments thereof. In a preferred embodiment the antibodies are human antigen binding antibody fragments of the present invention include, but are not limited to, Fab, Fab' F(ab)2 and F(ab')2, Fd, single-chain Fvs (scFv), single-chain antibodies, disulfide-linked Fvs (sdFv) and fragments comprising either a V_L or V_H domain. The antibodies may be from any animal origin including birds and mammals. Preferably, the antibodies are human, murine, rabbit, goat, guinea pig, camel, horse, or chicken.

Antigen-binding antibody fragments, including single-chain antibodies, may comprise the variable region(s) alone or in combination with the entire or partial of the following: hinge region, CH1, CH2, and CH3 domains. Also included in the invention are any combinations of variable region(s) and hinge region, CH1, CH2, and CH3 domains. The present invention further includes chimeric, humanized, and human monoclonal and polyclonal antibodies, which specifically bind the polypeptides of the present invention. The present invention further includes antibodies that are anti-idiotypic to the antibodies of the present invention.

The antibodies of the present invention may be monospecific, bispecific, and trispecific or have greater multispecificity. Multispecific antibodies may be specific for different epitopes of a polypeptide of the present invention or may be specific for both a polypeptide of the present invention as well as for heterologous compositions, such as a heterologous polypeptide or solid support material. *See, e.g.*, WO 93/17715; WO 92/08802; WO 91/00360; WO 92/05793; Tutt, *et al.* (1991); US Patents 5,573,920, 4,474,893, 5,601,819, 4,714,681, 4,925,648; Kostelny *et al.* (1992), which disclosures are hereby incorporated by reference in their entireties.

Antibodies of the present invention may be described or specified in terms of the epitope(s) or epitope-bearing portion(s) of a polypeptide of the present invention, which are recognized or specifically bound by the antibody. The antibodies may specifically bind a complete protein encoded by a nucleic acid of the present invention, or a fragment thereof, particularly, in the case of secreted proteins the mature protein or the signal peptide. Therefore, the epitope(s) or epitope bearing polypeptide portion(s) may be specified as described herein, *e.g.*, by N-terminal and C-terminal positions, by size in contiguous amino acid residues, or otherwise described herein (including the sequence listing). Antibodies which specifically bind any epitope or polypeptide of the present invention may also be excluded as individual species. Therefore, the present invention includes

antibodies that specifically bind specified polypeptides of the present invention, and allows for the exclusion of the same.

Thus, another embodiment of the present invention is a purified or isolated antibody capable of specifically binding to a polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID Nos: 242-482 and the sequences of the clone inserts of the deposited clone pool. In one aspect of this embodiment, the antibody is capable of binding to an epitope-containing polypeptide comprising at least 6 consecutive amino acids, preferably at least 8 to 10 consecutive amino acids, more preferably at least 12, 15, 20, 25, 30, 40, 50, or 100 consecutive amino acids of a sequence selected from the group consisting of SEQ ID Nos: 242-482 and sequences of the clone inserts of the deposited clone pool.

Antibodies of the present invention may also be described or specified in terms of their cross-reactivity. Antibodies that do not specifically bind any other analog, ortholog, or homologue of the polypeptides of the present invention are included. Antibodies that do not bind polypeptides with less than 95%, less than 90%, less than 85%, less than 80%, less than 75%, less than 70%, less than 65%, less than 60%, less than 55%, and less than 50% identity (as calculated using methods known in the art and described herein, e.g., using FASTDB and the parameters set forth herein) to a polypeptide of the present invention are also included in the present invention. Further included in the present invention are antibodies, which only bind polypeptides encoded by polynucleotides, which hybridize to a polynucleotide of the present invention under stringent hybridization conditions (as described herein). Antibodies of the present invention may also be described or specified in terms of their binding affinity. Preferred binding affinities include those with a dissociation constant or Kd less than $5X10^{-6}M$, $10^{-6}M$, $5X10^{-7}M$, $10^{-7}M$, $5X10^{-8}M$, $10^{-8}M$, $5X10^{-9}M$, $10^{-9}M$, $5X10^{-10}M$, $10^{-10}M$, $5X10^{-11}M$, $10^{-11}M$, $5X10^{-12}M$, $10^{-12}M$, $5X10^{-13}M$, $10^{-13}M$, $5X10^{-14}M$, $10^{-14}M$, $5X10^{-15}M$, and $10^{-15}M$.

The invention also concerns a purified or isolated antibody capable of specifically binding to a mutated GENSET protein or to a fragment or variant thereof comprising an epitope of the mutated GENSET protein.

Preparation of antibodies

The antibodies of the present invention may be prepared by any suitable method known in the art. Some of these methods are described in more detail in the example entitled "Preparation of Antibody Compositions to ". For example, a polypeptide of the present invention or an antigenic fragment thereof can be administered to an animal in order to induce the production of sera containing "polyclonal antibodies". As used herein, the term "monoclonal antibody" is not limited to antibodies produced through hybridoma technology but it rather refers to an antibody that is derived from a single clone, including eukaryotic, prokaryotic, or phage clone, and not the method

by which it is produced. Monoclonal antibodies can be prepared using a wide variety of techniques known in the art including the use of hybridoma, recombinant, and phage display technology.

Hybridoma techniques include those known in the art (*See, e.g.,* Harlow *et al.* 1988; Hammerling, *et al,* 1981). (Said references incorporated by reference in their entireties). Fab and F(ab')2 fragments may be produced, for example, from hybridoma-produced antibodies by proteolytic cleavage, using enzymes such as papain (to produce Fab fragments) or pepsin (to produce F(ab')2 fragments).

Alternatively, antibodies of the present invention can be produced through the application of recombinant DNA technology or through synthetic chemistry using methods known in the art. For example, the antibodies of the present invention can be prepared using various phage display methods known in the art. In phage display methods, functional antibody domains are displayed on the surface of a phage particle, which carries polynucleotide sequences encoding them. Phage with a desired binding property are selected from a repertoire or combinatorial antibody library (e.g. human or murine) by selecting directly with antigen, typically antigen bound or captured to a solid surface or bead. Phage used in these methods are typically filamentous phage including fd and M13 with Fab, Fv or disulfide stabilized Fv antibody domains recombinantly fused to either the phage gene III or gene VIII protein. Examples of phage display methods that can be used to make the antibodies of the present invention include those disclosed in Brinkman *et al.* (1995); Ames, *et al.* (1995); Kettleborough, *et al.* (1994); Persic, *et al.* (1997); Burton *et al.* (1994); PCT/GB91/01134; WO 90/02809; WO 91/10737; WO 92/01047; WO 92/18619; WO 93/11236; WO 95/15982; WO 95/20401; and US Patents 5,698,426, 5,223,409, 5,403,484, 5,580,717, 5,427,908, 5,750,753, 5,821,047, 5,571,698, 5,427,908, 5,516,637, 5,780,225, 5,658,727 and 5,733,743 (said references incorporated by reference in their entireties).

As described in the above references, after phage selection, the antibody coding regions from the phage can be isolated and used to generate whole antibodies, including human antibodies, or any other desired antigen binding fragment, and expressed in any desired host including mammalian cells, insect cells, plant cells, yeast, and bacteria. For example, techniques to recombinantly produce Fab, Fab' F(ab)2 and F(ab')2 fragments can also be employed using methods known in the art such as those disclosed in WO 92/22324; Mullinax *et al.* (1992); and Sawai *et al.* (1995); and Better *et al.* (1988) (said references incorporated by reference in their entireties).

Examples of techniques which can be used to produce single-chain Fvs and antibodies include those described in U.S. Patents 4,946,778 and 5,258,498; Huston *et al.* (1991); Shu *et al.* (1993); and Skerra *et al.* (1988), which disclosures are hereby incorporated by reference in their entireties. For some uses, including *in vivo* use of antibodies in humans and *in vitro* detection assays, it may be preferable to use chimeric, humanized, or human antibodies. Methods for producing chimeric antibodies are known in the art. *See e.g.*, Morrison, (1985); Oi *et al.*, (1986);

Gillies *et al.* (1989); and US Patent 5,807,715, which disclosures are hereby incorporated by reference in their entireties. Antibodies can be humanized using a variety of techniques including CDR-grafting (EP 0 239 400; WO 91/09967; US Patent 5,530,101; and 5,585,089), veneering or resurfacing, (EP 0 592 106; EP 0 519 596; Padlan, 1991; Studnicka *et al.*, 1994; Roguska *et al.*, 1994), and chain shuffling (US Patent 5,565,332), which disclosures are hereby incorporated by reference in their entireties. Human antibodies can be made by a variety of methods known in the art including phage display methods described above. *See also*, US Patents 4,444,887, 4,716,111, 5,545,806, and 5,814,318; WO 98/46645; WO 98/50433; WO 98/24893; WO 96/34096; WO 96/33735; and WO 91/10741 (said references incorporated by reference in their entireties).

Further included in the present invention are antibodies recombinantly fused or chemically conjugated (including both covalently and non-covalently conjugations) to a polypeptide of the present invention. The antibodies may be specific for antigens other than polypeptides of the present invention. For example, antibodies of the present invention may be recombinantly fused or conjugated to molecules useful as labels in detection assays and effector molecules such as heterologous polypeptides, drugs, or toxins. See, e.g., WO 92/08495; WO 91/14438; WO 89/12624; US Patent 5,314,995; and EP 0 396 387, which disclosures are hereby incorporated by reference in their entireties. Fused antibodies may also be used to target the polypeptides of the present invention to particular cell types, either *in vitro* or *in vivo,* by fusing or conjugating the polypeptides of the present invention to antibodies specific for particular cell surface receptors. Antibodies fused or conjugated to the polypeptides of the present invention may also be used in *vitro* immunoassays and purification methods using methods known in the art (*See e.g.*, Harbor *et al. supra*; WO 93/21232; EP 0 439 095; Naramura, M. *et al.* 1994; US Patent 5,474,981; Gillies *et al.*, 1992; Fell *et al.*, 1991) (said references incorporated by reference in their entireties).

The present invention further includes compositions comprising the polypeptides of the present invention fused or conjugated to antibody domains other than the variable regions. For example, the polypeptides of the present invention may be fused or conjugated to an antibody Fc region, or portion thereof. The antibody portion fused to a polypeptide of the present invention may comprise the hinge region, CH1 domain, CH2 domain, and CH3 domain or any combination of whole domains or portions thereof. The polypeptides of the present invention may be fused or conjugated to the above antibody portions to increase the *in vivo* half-life of the polypeptides or for use in immunoassays using methods known in the art. The polypeptides may also be fused or conjugated to the above antibody portions to form multimers. For example, Fc portions fused to the polypeptides of the present invention can form dimers through disulfide bonding between the Fc portions. Higher multimeric forms can be made by fusing the polypeptides to portions of IgA and IgM. Methods for fusing or conjugating the polypeptides of the present invention to antibody portions are known in the art. *See e.g.*, US Patents 5,336,603, 5,622,929, 5,359,046, 5,349,053, 5,447,851, 5,112,946; EP 0 307 434, EP 0 367 166; WO 96/04388, WO 91/06570; Ashkenazi *et al.*

(1991); Zheng *et al.* (1995); and Vil *et al.* (1992) (said references incorporated by reference in their entireties).

Non-human animals or mammals, whether wild-type or transgenic, which express a different species of GENSET than the one to which antibody binding is desired, and animals which do not express GENSET (i.e. a GENSET knock out animal as described herein) are particularly useful for preparing antibodies. GENSET knock out animals will recognize all or most of the exposed regions of a GENSET protein as foreign antigens, and therefore produce antibodies with a wider array of GENSET epitopes. Moreover, smaller polypeptides with only 10 to 30 amino acids may be useful in obtaining specific binding to any one of the GENSET proteins . In addition, the humoral immune system of animals which produce a species of GENSET that resembles the antigenic sequence will preferentially recognize the differences between the animal's native GENSET species and the antigen sequence, and produce antibodies to these unique sites in the antigen sequence. Such a technique will be particularly useful in obtaining antibodies that specifically bind to any one of the GENSET proteins .

The antibodies of the invention may be labeled by any one of the radioactive, fluorescent or enzymatic labels known in the art.

USES OF POLYNUCLEOTIDES

Uses of polynucleotides as reagents

The polynucleotides of the present invention, particularly those described in the "Oligonucleotide primers and probes" section, may be used as reagents in isolation procedures, diagnostic assays, and forensic procedures. For example, sequences from the GENSET polynucleotides of the invention may be detectably labeled and used as probes to isolate other sequences capable of hybridizing to them. In addition, sequences from the GENSET polynucleotides of the invention may be used to design PCR primers to be used in isolation, diagnostic, or forensic procedures.

In forensic analyses

PCR primers may be used in forensic analyses, such as the DNA fingerprinting techniques described below. Such analyses may utilize detectable probes or primers based on the sequences of the polynucleotides of the invention. Consequently, the present invention encompasses methods of identification of an individual using the polynucleotides of the invention in forensic analyses, wherein said method includes the steps of:

a) obtaining a biological sample containing nucleic acid material from an individual;

b) obtaining an identification pattern for this individual using the polynucleotides of the invention, particularly using GENSET primers and probes;

c) comparing said identification pattern with a reference identification pattern; and

d) determining whether said identification pattern is identical to said reference identification pattern.

In one embodiment of this method, the identification pattern consists in sequences of amplicons obtained using GENSET primers as explained in the sections entitled "Forensic Matching by DNA Sequencing" and "Positive Identification by DNA Sequencing".

In another embodiment, the identification pattern consists in unique band or dot patterns obtained using any method described in the sections entitled "Southern Blot Forensic Identification", "Dot Blot Identification Procedure" and "Alternative "Fingerprint" Identification Technique".

Forensic Matching by DNA Sequencing

In one exemplary method, DNA samples are isolated from forensic specimens of, for example, hair, semen, blood or skin cells by conventional methods. A panel of PCR primers designed from different polynucleotides of the invention using any technique known to those skilled in the art including those described herein, is then utilized to amplify DNA of approximately 100-200 bases in length from the forensic specimen. Corresponding sequences are obtained from a test subject. Each of these identification DNAs is then sequenced using standard techniques, and a simple database comparison determines the differences, if any, between the sequences from the subject and those from the sample. Statistically significant differences between the suspect's DNA sequences and those from the sample conclusively prove a lack of identity. This lack of identity can be proven, for example, with only one sequence. Identity, on the other hand, should be demonstrated with a large number of sequences, all matching. Preferably, a minimum of 50 statistically identical sequences of 100 bases in length are used to prove identity between the suspect and the sample.

Positive Identification by DNA Sequencing

The "Forensic Matching by DNA Sequencing" technique described herein may also be used on a larger scale to provide a unique fingerprint-type identification of any individual. In this technique, primers are prepared from a large number of polynucleotides of the invention. Preferably, 20 to 50 different primers are used. These primers are used to obtain a corresponding number of PCR-generated DNA segments from the individual in question. Each of these DNA segments is sequenced. The database of sequences generated through this procedure uniquely identifies the individual from whom the sequences were obtained. The same panel of primers may then be used at any later time to absolutely correlate tissue or other biological specimen with that individual.

Southern Blot Forensic Identification

The "Positive Identification by DNA Sequencing" procedure described herein is repeated to obtain a panel of at least 10 amplified sequences from an individual and a specimen. Preferably, the panel contains at least 50 amplified sequences. More preferably, the panel contains 100 amplified sequences. In some embodiments, the panel contains 200 amplified sequences. This PCR-generated DNA is then digested with one or a combination of, preferably, four base specific restriction enzymes. Such enzymes are commercially available and known to those of skill in the art. After digestion, the resultant gene fragments are size separated in multiple duplicate wells on an agarose gel and transferred to nitrocellulose using Southern blotting techniques well known to those with skill in the art. For a review of Southern blotting see Davis *et al.* (1986), which disclosure is hereby incorporated by reference in its entirety.

A panel of probes based on the sequences of the polynucleotides of the invention, or fragments thereof of at least 10 bases, are radioactively or colorimetrically labeled using methods known in the art, such as nick translation or end labeling, and hybridized to the Southern blot using techniques known in the art. Preferably, the probe comprises at least 12, 15, or 17 consecutive nucleotides from the polynucleotide of the invention. More preferably, the probe comprises at least 20-30 consecutive nucleotides from the polynucleotide of the invention. In some embodiments, the probe comprises more than 30 nucleotides from the polynucleotide of the invention. In other embodiments, the probe comprises at least 40, at least 50, at least 75, at least 100, at least 150, or at least 200 consecutive nucleotides from the polynucleotide of the invention.

Preferably, at least 5 to 10 of these labeled probes are used, and more preferably at least about 20 or 30 are used to provide a unique pattern. The resultant bands appearing from the hybridization of a large sample of polynucleotide of the invention will be a unique identifier. Since the restriction enzyme cleavage will be different for every individual, the band pattern on the Southern blot will also be unique. Increasing the number of cDNA probes will provide a statistically higher level of confidence in the identification since there will be an increased number of sets of bands used for identification.

Dot Blot Identification Procedure

Another technique for identifying individuals using the polynucleotide sequences disclosed herein utilizes a dot blot hybridization technique.

Genomic DNA is isolated from nuclei of subject to be identified. Oligonucleotide probes of approximately 30 bp in length are synthesized that correspond to at least 10, preferably 50 sequences from the polynucleotide of the invention. The probes are used to hybridize to the genomic DNA through conditions known to those in the art. The oligonucleotides are end labeled with P^{32} using polynucleotide kinase (Pharmacia). Dot Blots are created by spotting the genomic DNA onto nitrocellulose or the like using a vacuum dot blot manifold (BioRad, Richmond California). The nitrocellulose filter containing the genomic sequences is baked or UV linked to the

filter, prehybridized and hybridized with labeled probe using techniques known in the art (Davis *et al.* 1986). The ^{32}P labeled DNA fragments are sequentially hybridized with successively stringent conditions to detect minimal differences between the 30 bp sequence and the DNA. Tetramethylammonium chloride is useful for identifying clones containing small numbers of nucleotide mismatches (Wood *et al.*, 1985). A unique pattern of dots distinguishes one individual from another individual.

Alternative "Fingerprint" Identification Technique

In a representative alternative fingerprinting procedure, the probes are derived from cDNAs. Preferably, a plurality of probes having sequences from different genes are used as follows. Polynucleotides containing at least 10 consecutive bases from these sequences can be used as probes. Preferably, the probe comprises at least 12, 15, or 17 consecutive nucleotides from the polynucleotide of the invention. More preferably, the probe comprises at least 20-30 consecutive nucleotides from the polynucleotide of the invention. In some embodiments, the probe comprises more than 30 nucleotides from the polynucleotide of the invention. In other embodiments, the probe comprises at least 40, at least 50, at least 75, at least 100, at least 150, or at least 200 consecutive nucleotides from the polynucleotide of the invention.

Oligonucleotides, generally 20-mers, are prepared from a large number, e.g. 50, 100, or 200, of polynucleotides of the invention using commercially available oligonucleotide services such as Genset, Paris, France. Cell samples from the test subject are processed for DNA using techniques well known to those with skill in the art. The nucleic acid is digested with restriction enzymes such as EcoRI and XbaI. Following digestion, samples are applied to wells for electrophoresis. The procedure, as known in the art, may be modified to accommodate polyacrylamide electrophoresis, however in this example, samples containing 5 ug of DNA are loaded into wells and separated on 0.8% agarose gels. The gels are transferred onto nitrocellulose using standard Southern blotting techniques.

10 ng of each of the oligonucleotides are pooled and end-labeled with P^{32}. The nitrocellulose is prehybridized with blocking solution and hybridized with the labeled probes. Following hybridization and washing, the nitrocellulose filter is exposed to X-Omat AR X-ray film. The resulting hybridization pattern will be unique for each individual.

It is additionally contemplated within this example that the number of probe sequences used can be varied for additional accuracy or clarity.

To find corresponding genomic DNA sequences

The GENSET cDNAs of the invention may also be used to clone sequences located upstream of the cDNAs of the invention on the corresponding genomic DNA. Such upstream sequences may be capable of regulating gene expression, including promoter sequences, enhancer sequences, and other upstream sequences which influence transcription or translation levels. Once

identified and cloned, these upstream regulatory sequences may be used in expression vectors designed to direct the expression of an inserted gene in a desired spatial, temporal, developmental, or quantitative fashion.

Use of cDNAs or Fragments thereof to Clone Upstream Sequences from Genomic DNA

Sequences derived from polynucleotides of the inventions may be used to isolate the promoters of the corresponding genes using chromosome walking techniques. In one chromosome walking technique, which utilizes the GenomeWalker™ kit available from Clontech, five complete genomic DNA samples are each digested with a different restriction enzyme which has a 6 base recognition site and leaves a blunt end. Following digestion, oligonucleotide adapters are ligated to each end of the resulting genomic DNA fragments.

For each of the five genomic DNA libraries, a first PCR reaction is performed according to the manufacturer's instructions (which are incorporated herein by reference) using an outer adaptor primer provided in the kit and an outer gene specific primer. The gene specific primer should be selected to be specific for the polynucleotide of the invention of interest and should have a melting temperature, length, and location in the polynucleotide of the invention which is consistent with its use in PCR reactions. Each first PCR reaction contains 5ng of genomic DNA, 5 µl of 10X Tth reaction buffer, 0.2 mM of each dNTP, 0.2 µM each of outer adaptor primer and outer gene specific primer, 1.1 mM of $Mg(OAc)_2$, and 1 µl of the Tth polymerase 50X mix in a total volume of 50 µl. The reaction cycle for the first PCR reaction is as follows: 1 min at 94 degree Celsius / 2 sec at 94 degree Celsius, 3 min at 72 degree Celsius (7 cycles) / 2 sec at 94 degree Celsius, 3 min at 67 degree Celsius (32 cycles) / 5 min at 67 degree Celsius.

The product of the first PCR reaction is diluted and used as a template for a second PCR reaction according to the manufacturer's instructions using a pair of nested primers which are located internally on the amplicon resulting from the first PCR reaction. For example, 5 µl of the reaction product of the first PCR reaction mixture may be diluted 180 times. Reactions are made in a 50 µl volume having a composition identical to that of the first PCR reaction except the nested primers are used. The first nested primer is specific for the adaptor, and is provided with the GenomeWalker™ kit. The second nested primer is specific for the particular polynucleotide of the invention for which the promoter is to be cloned and should have a melting temperature, length, and location in the polynucleotide of the invention which is consistent with its use in PCR reactions. The reaction parameters of the second PCR reaction are as follows: 1 min at 94 degree Celsius / 2 sec at 94 degree Celsius, 3 min at 72 degree Celsius (6 cycles) / 2 sec at 94 degree Celsius, 3 min at 67 degree Celsius (25 cycles) / 5 min at 67 degree Celsius

The product of the second PCR reaction is purified, cloned, and sequenced using standard techniques. Alternatively, two or more human genomic DNA libraries can be constructed by using two or more restriction enzymes. The digested genomic DNA is cloned into vectors which can be

converted into single stranded, circular, or linear DNA. A biotinylated oligonucleotide comprising at least 15 nucleotides from the polynucleotide of the invention sequence is hybridized to the single stranded DNA. Hybrids between the biotinylated oligonucleotide and the single stranded DNA containing the polynucleotide of the invention sequence are isolated as described herein. Thereafter, the single stranded DNA containing the polynucleotide of the invention sequence is released from the beads and converted into double stranded DNA using a primer specific for the polynucleotide of the invention sequence or a primer corresponding to a sequence included in the cloning vector. The resulting double stranded DNA is transformed into bacteria. DNAs containing the GENSET polynucleotide sequences are identified by colony PCR or colony hybridization.

Identification of Promoters in Cloned Upstream Sequences

Once the upstream genomic sequences have been cloned and sequenced as described above, prospective promoters and transcription start sites within the upstream sequences may be identified by comparing the sequences upstream of the polynucleotides of the inventions with databases containing known transcription start sites, transcription factor binding sites, or promoter sequences.

In addition, promoters in the upstream sequences may be identified using promoter reporter vectors as follows. The expression of the reporter gene will be detected when placed under the control of regulatory active polynucleotide fragments or variants of the GENSET promoter region located upstream of the first exon of the GENSET gene. Suitable promoter reporter vectors, into which the GENSET promoter sequences may be cloned include pSEAP-Basic, pSEAP-Enhancer, pβgal-Basic, pβgal-Enhancer, or pEGFP-1 Promoter Reporter vectors available from Clontech, or pGL2-basic or pGL3-basic promoterless luciferase reporter gene vector from Promega. Briefly, each of these promoter reporter vectors include multiple cloning sites positioned upstream of a reporter gene encoding a readily assayable protein such as secreted alkaline phosphatase, luciferase, beta-galactosidase, or green fluorescent protein. The sequences upstream the GENSET coding region are inserted into the cloning sites upstream of the reporter gene in both orientations and introduced into an appropriate host cell. The level of reporter protein is assayed and compared to the level obtained from a vector which lacks an insert in the cloning site. The presence of an elevated expression level in the vector containing the insert with respect to the control vector indicates the presence of a promoter in the insert. If necessary, the upstream sequences can be cloned into vectors which contain an enhancer for increasing transcription levels from weak promoter sequences. A significant level of expression above that observed with the vector lacking an insert indicates that a promoter sequence is present in the inserted upstream sequence.

Promoter sequence within the upstream genomic DNA may be further defined by site directed mutagenesis, linker scanning analysis, or other techniques familiar to those skilled in the art. For example, the boundaries of promoters may be further investigated by constructing nested 5' and/or 3' deletions in the upstream DNA using conventional techniques such as Exonuclease III or

appropriate restriction endonuclease digestion. The resulting deletion fragments can be inserted into the promoter reporter vector to determine whether the deletion has increased, reduced or illuminated promoter activity, such as described, for example, by Coles *et al.* (1998), the disclosure of which is incorporated herein by reference in its entirety. In this way, the boundaries of the promoters may be defined. If desired, potential individual regulatory sites within the promoter may be identified using site directed mutagenesis or linker scanning to obliterate potential transcription factor binding sites within the promoter individually or in combination. The effects of these mutations on transcription levels may be determined by inserting the mutations into cloning sites in promoter reporter vectors. This type of assay is well known to those skilled in the art and is described in WO 97/17359, US Patent No. 5,374,544; EP 582 796; US Patent No. 5,698,389; US 5,643,746; US Patent No. 5,502,176; and US Patent 5,266,488; the disclosures of which are incorporated by reference herein in their entirety.

The strength and the specificity of the promoter of each GENSET gene can be assessed through the expression levels of a detectable polynucleotide operably linked to the GENSET promoter in different types of cells and tissues. The detectable polynucleotide may be either a polynucleotide that specifically hybridizes with a predefined oligonucleotide probe, or a polynucleotide encoding a detectable protein, including a GENSET polypeptide or a fragment or a variant thereof. This type of assay is well known to those skilled in the art and is described in US Patent No. 5,502,176; and US Patent No. 5,266,488; the disclosures of which are incorporated by reference herein in their entirety. Some of the methods are discussed in more detail elsewhere in the application.

The promoters and other regulatory sequences located upstream of the polynucleotides of the inventions may be used to design expression vectors capable of directing the expression of an inserted gene in a desired spatial, temporal, developmental, or quantitative manner. A promoter capable of directing the desired spatial, temporal, developmental, and quantitative patterns may be selected using the results of the expression analysis described herein. For example, if a promoter which confers a high level of expression in muscle is desired, the promoter sequence upstream of a polynucleotide of the invention derived from an mRNA which is expressed at a high level in muscle may be used in the expression vector. Such vectors are described in more detail elsewhere in the application.

Preferably, the desired promoter is placed near multiple restriction sites to facilitate the cloning of the desired insert downstream of the promoter, such that the promoter is able to drive expression of the inserted gene. The promoter may be inserted in conventional nucleic acid backbones designed for extrachromosomal replication, integration into the host chromosomes or transient expression. Suitable backbones for the present expression vectors include retroviral backbones, backbones from eukaryotic episomes such as SV40 or Bovine Papilloma Virus, backbones from bacterial episomes, or artificial chromosomes.

Preferably, the expression vectors also include a polyA signal downstream of the multiple restriction sites for directing the polyadenylation of mRNA transcribed from the gene inserted into the expression vector.

To find similar sequences

Polynucleotides of the invention may be used to isolate and/or purify nucleic acids similar thereto using any methods well known to those skilled in the art including the techniques based on hybridization or on amplification described in this section. These methods may be used to obtain the genomic DNAs which encode the mRNAs from which the GENSET cDNAs are derived, mRNAs corresponding to GENSET cDNAs, or nucleic acids which are homologous to GENSET cDNAs or fragments thereof, such as variants, species homologues or orthologs. Thus, a plurality of cDNAs similar to GENSET polynucleotides may be provided as cDNA libraries for subsequent evaluation of the encoded proteins or use in diagnostic assays as described herein. cDNAs prepared by any method described therein may be subsequently engineered to obtain nucleic acids which include desired fragments of the cDNA using conventional techniques such as subcloning, PCR, or *in vitro* oligonucleotide synthesis. For example, nucleic acids which include only the coding sequences may be obtained using techniques known to those skilled in the art. Similarly, nucleic acids containing any other desired fragment of the coding sequences for the encoded protein may be obtained.

Indeed, cDNAs of the present invention or fragments thereof may be used to isolate nucleic acids similar to cDNAs from a cDNA library or a genomic DNA library. Such cDNA libraries or genomic DNA libraries may be obtained from a commercial source or made using techniques familiar to those skilled in the art such as those described in PCT publication WO 00/37491, which disclosure is hereby incorporated by reference in its entirety. Examples of methods for obtaining nucleic acids similar to GENSET polynucleotides are described below.

Hybridization-based methods

Techniques for identifying cDNA clones in a cDNA library which hybridize to a given probe sequence are disclosed in Sambrook *et al.*, (1989) and in Hames and Higgins (1985), the disclosures of which are incorporated herein by reference in their entireties. The same techniques may be used to isolate genomic DNAs.

Briefly, cDNA or genomic DNA clones which hybridize to the detectable probe are identified and isolated for further manipulation as follows. Any polynucleotide fragment of the invention may be used as a probe, in particular those defined in the "Oligonucleotide primers and probes" section. A probe comprising at least 10 consecutive nucleotides from a GENSET cDNA or fragment thereof is labeled with a detectable label such as a radioisotope or a fluorescent molecule. Preferably, the probe comprises at least 12, 15, or 17 consecutive nucleotides from the cDNA or fragment thereof. More preferably, the probe comprises 20 to 30 consecutive nucleotides from the

cDNA or fragment thereof. In some embodiments, the probe comprises more than 30 nucleotides from the cDNA or fragment thereof.

Techniques for labeling the probe are well known and include phosphorylation with polynucleotide kinase, nick translation, *in vitro* transcription, and non radioactive techniques. The cDNAs or genomic DNAs in the library are transferred to a nitrocellulose or nylon filter and denatured. After blocking of non specific sites, the filter is incubated with the labeled probe for an amount of time sufficient to allow binding of the probe to cDNAs or genomic DNAs containing a sequence capable of hybridizing thereto.

By varying the stringency of the hybridization conditions used to identify cDNAs or genomic DNAs which hybridize to the detectable probe, cDNAs or genomic DNAs having different levels of identity to the probe can be identified and isolated as described below.

Stringent conditions

"Stringent hybridization conditions" are defined as conditions in which only nucleic acids having a high level of identity to the probe are able to hybridize to said probe. These conditions may be calculated as follows:

For probes between 14 and 70 nucleotides in length the melting temperature (Tm) is calculated using the formula: Tm=81.5+16.6(log (Na+))+0.41(fraction G+C)-(600/N) where N is the length of the probe.

If the hybridization is carried out in a solution containing formamide, the melting temperature may be calculated using the equation: Tm=81.5+16.6(log (Na+))+0.41(fraction G+C)-(0.63% formamide)-(600/N) where N is the length of the probe.

Prehybridization may be carried out in 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100 μg denatured fragmented salmon sperm DNA or 6X SSC, 5X Denhardt's reagent, 0.5% SDS, 100 μg denatured fragmented salmon sperm DNA, 50% formamide. The formulas for SSC and Denhardt's solutions are listed in Sambrook *et al.*, 1986.

Hybridization is conducted by adding the detectable probe to the prehybridization solutions listed above. Where the probe comprises double stranded DNA, it is denatured before addition to the hybridization solution. The filter is contacted with the hybridization solution for a sufficient period of time to allow the probe to hybridize to nucleic acids containing sequences complementary thereto or homologous thereto. For probes over 200 nucleotides in length, the hybridization may be carried out at 15-25°C below the Tm. For shorter probes, such as oligonucleotide probes, the hybridization may be conducted at 15-25°C below the Tm. Preferably, for hybridizations in 6X SSC, the hybridization is conducted at approximately 68°C. Preferably, for hybridizations in 50% formamide containing solutions, the hybridization is conducted at approximately 42°C.

Following hybridization, the filter is washed in 2X SSC, 0.1% SDS at room temperature for 15 minutes. The filter is then washed with 0.1X SSC, 0.5% SDS at room temperature for 30

minutes to 1 hour. Thereafter, the solution is washed at the hybridization temperature in 0.1X SSC, 0.5% SDS. A final wash is conducted in 0.1X SSC at room temperature.

Nucleic acids which have hybridized to the probe are identified by autoradiography or other conventional techniques.

Low and moderate conditions

Changes in the stringency of hybridization and signal detection are primarily accomplished through the manipulation of formamide concentration (lower percentages of formamide result in lowered stringency); salt conditions, or temperature. The above procedure may thus be modified to identify nucleic acids having decreasing levels of identity to the probe sequence. For example, the hybridization temperature may be decreased in increments of 5°C from 68°C to 42°C in a hybridization buffer having a sodium concentration of approximately 1M. Following hybridization, the filter may be washed with 2X SSC, 0.5% SDS at the temperature of hybridization. These conditions are considered to be "moderate" conditions above 50°C and "low" conditions below 50°C. Alternatively, the hybridization may be carried out in buffers, such as 6X SSC, containing formamide at a temperature of 42°C. In this case, the concentration of formamide in the hybridization buffer may be reduced in 5% increments from 50% to 0% to identify clones having decreasing levels of identity to the probe. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at 50°C. These conditions are considered to be "moderate" conditions above 25% formamide and "low" conditions below 25% formamide. cDNAs or genomic DNAs which have hybridized to the probe are identified by autoradiography or other conventional techniques.

Note that variations in the above conditions may be accomplished through the inclusion and/or substitution of alternate blocking reagents used to suppress background in hybridization experiments. Typical blocking reagents include Denhardt's reagent, BLOTTO, heparin, denatured salmon sperm DNA, and commercially available proprietary formulations. The inclusion of specific blocking reagents may require modification of the hybridization conditions described above, due to problems with compatibility.

Consequently, the present invention encompasses methods of isolating nucleic acids similar to the polynucleotides of the invention, comprising the steps of:

a) contacting a collection of cDNA or genomic DNA molecules with a detectable probe comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40 or 50 consecutive nucleotides of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 1-241, the sequences of clones inserts of the deposited clone pool and sequences complementary thereto under stringent, moderate or low conditions which permit said probe to hybridize to at least a cDNA or genomic DNA molecule in said collection;

b) identifying said cDNA or genomic DNA molecule which hybridizes to said detectable probe; and

c) isolating said cDNA or genomic DNA molecule which hybridized to said probe.

PCR-based methods

In addition to the above described methods, other protocols are available to obtain homologous cDNAs using GENSET cDNA of the present invention or fragment thereof as outlined in the following paragraphs.

cDNAs may be prepared by obtaining mRNA from the tissue, cell, or organism of interest using mRNA preparation procedures utilizing polyA selection procedures or other techniques known to those skilled in the art. A first primer capable of hybridizing to the polyA tail of the mRNA is hybridized to the mRNA and a reverse transcription reaction is performed to generate a first cDNA strand.

The term "capable of hybridizing to the polyA tail of said mRNA" refers to and embraces all primers containing stretches of thymidine residues, so-called oligo(dT) primers, that hybridize to the 3' end of eukaryotic poly(A)+ mRNAs to prime the synthesis of a first cDNA strand. Techniques for generating said oligo (dT) primers and hybridizing them to mRNA to subsequently prime the reverse transcription of said hybridized mRNA to generate a first cDNA strand are well known to those skilled in the art and are described in Current Protocols in Molecular Biology, John Wiley and Sons, Inc. 1997 and Sambrook *et al.*, 1989. Preferably, said oligo (dT) primers are present in a large excess in order to allow the hybridization of all mRNA 3'ends to at least one oligo (dT) molecule. The priming and reverse transcription steps are preferably performed between 37°C and 55°C depending on the type of reverse transcriptase used. Preferred oligo(dT) primers for priming reverse transcription of mRNAs are oligonucleotides containing a stretch of thymidine residues of sufficient length to hybridize specifically to the polyA tail of mRNAs, preferably of 12 to 18 thymidine residues in length. More preferably, such oligo(T) primers comprise an additional sequence upstream of the poly(dT) stretch in order to allow the addition of a given sequence to the 5'end of all first cDNA strands which may then be used to facilitate subsequent manipulation of the cDNA. Preferably, this added sequence is 8 to 60 residues in length. For instance, the addition of a restriction site in 5' of cDNAs facilitates subcloning of the obtained cDNA. Alternatively, such an added 5'end may also be used to design primers of PCR to specifically amplify cDNA clones of interest.

The first cDNA strand is then hybridized to a second primer. Any polynucleotide fragment of the invention may be used, and in particular those described in the "Oligonucleotide primers and probes" section. This second primer contains at least 10 consecutive nucleotides of a polynucleotide of the invention. Preferably, the primer comprises at least 10, 12, 15, 17, 18, 20, 23, 25, or 28 consecutive nucleotides of a polynucleotide of the invention. In some embodiments, the primer comprises more than 30 nucleotides of a polynucleotide of the invention. If it is desired to obtain cDNAs containing the full protein coding sequence, including the authentic translation initiation

site, the second primer used contains sequences located upstream of the translation initiation site. The second primer is extended to generate a second cDNA strand complementary to the first cDNA strand. Alternatively, RT-PCR may be performed as described above using primers from both ends of the cDNA to be obtained.

The double stranded cDNAs made using the methods described above are isolated and cloned. The cDNAs may be cloned into vectors such as plasmids or viral vectors capable of replicating in an appropriate host cell. For example, the host cell may be a bacterial, mammalian, avian, or insect cell.

Techniques for isolating mRNA, reverse transcribing a primer hybridized to mRNA to generate a first cDNA strand, extending a primer to make a second cDNA strand complementary to the first cDNA strand, isolating the double stranded cDNA and cloning the double stranded cDNA are well known to those skilled in the art and are described *in Current Protocols in Molecular Biology*, John Wiley & Sons, Inc. 1997 and Sambrook *et al.*, 1989.

Consequently, the present invention encompasses methods of making cDNAs. In a first embodiment, the method of making a cDNA comprises the steps of

a) contacting a collection of mRNA molecules from human cells with a primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of the sequences complementary to SEQ ID Nos: 1-241 and sequences complementary to a clone insert of the deposited clone pool;

b) hybridizing said primer to an mRNA in said collection;

c) reverse transcribing said hybridized primer to make a first cDNA strand from said mRNA;

d) making a second cDNA strand complementary to said first cDNA strand; and

e) isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

Another embodiment of the present invention is a purified cDNA obtainable by the method of the preceding paragraph. In one aspect of this embodiment, the cDNA encodes at least a portion of a human polypeptide.

In a second embodiment, the method of making a cDNA comprises the steps of

a) contacting a collection of mRNA molecules from human cells with a first primer capable of hybridizing to the polyA tail of said mRNA;

b) hybridizing said first primer to said polyA tail;

c) reverse transcribing said mRNA to make a first cDNA strand;

d) making a second cDNA strand complementary to said first cDNA strand using at least one primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool; and

e) isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

In another aspect of this method the second cDNA strand is made by

a) contacting said first cDNA strand with a second primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, and a third primer which sequence is fully included within the sequence of said first primer;

b) performing a first polymerase chain reaction with said second and third primers to generate a first PCR product;

c) contacting said first PCR product with a fourth primer, comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of said sequence selected from the group consisting of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, and a fifth primer, which sequence is fully included within the sequence of said third primer, wherein said fourth and fifth hybridize to sequences within said first PCR product; and

d) performing a second polymerase chain reaction, thereby generating a second PCR product.

Alternatively, the second cDNA strand may be made by contacting said first cDNA strand with a second primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, and a third primer which sequence is fully included within the sequence of said first primer and performing a polymerase chain reaction with said second and third primers to generate said second cDNA strand.

Alternatively, the second cDNA strand may be made by

a) contacting said first cDNA strand with a second primer comprising at least 12, 15, 18, 20, 23, 25, 28, 30, 35, 40, or 50 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool;

b) hybridizing said second primer to said first strand cDNA; and

c) extending said hybridized second primer to generate said second cDNA strand.

Another embodiment of the present invention is a purified cDNA obtainable by a method of making a cDNA of the invention. In one aspect of this embodiment, said cDNA encodes at least a portion of a human polypeptide.

Other protocols

Alternatively, other procedures may be used for obtaining homologous cDNAs. In one approach, cDNAs are prepared from mRNA and cloned into double stranded phagemids as follows. The cDNA library in the double stranded phagemids is then rendered single stranded by treatment with an endonuclease, such as the Gene II product of the phage F1 and an exonuclease (Chang *et*

al., 1993, which disclosure is hereby incorporated by reference in its entirety). A biotinylated oligonucleotide comprising the sequence of a fragment of a known GENSET cDNA, genomic DNA or fragment thereof is hybridized to the single stranded phagemids. Preferably, the fragment comprises at least 10, 12, 15, 17, 18, 20, 23, 25, or 28 consecutive nucleotides of a sequence selected from the group consisting of the sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool.

Hybrids between the biotinylated oligonucleotide and phagemids are isolated by incubating the hybrids with streptavidin coated paramagnetic beads and retrieving the beads with a magnet (Fry *et al,* 1992, which disclosure is hereby incorporated by reference in its entirety). Thereafter, the resulting phagemids are released from the beads and converted into double stranded DNA using a primer specific for the GENSET cDNA or fragment used to design the biotinylated oligonucleotide. Alternatively, protocols such as the Gene Trapper kit (Gibco BRL), which disclosure is which disclosure is hereby incorporated by reference in its entirety, may be used. The resulting double stranded DNA is transformed into bacteria. Homologous cDNAs to the GENSET cDNA or fragment thereof sequence are identified by colony PCR or colony hybridization.

As a chromosome marker

Chromosomal localization of the cDNA of the present invention were determined using information from public and proprietary databases. Table VIII lists the putative chromosomal location of the polynucleotides of the present invention. Column one lists the sequence identification number with the corresponding chromosomal location listed in column two. Thus, the present invention also relates to methods and compositions using the chromosomal location of the polynucleotides of the invention to construct a human high resolution map or to identify a given chromosome in a sample using any techniques known to those skilled in the art including those disclosed below.

GENSET polynucleotides may also be mapped to their chromosomal locations using any methods or techniques known to those skilled in the art including radiation hybrid (RH) mapping, PCR-based mapping and Fluorescence in situ hybridization (FISH) mapping described below.

Radiation hybrid mapping

Radiation hybrid (RH) mapping is a somatic cell genetic approach that can be used for high resolution mapping of the human genome. In this approach, cell lines containing one or more human chromosomes are lethally irradiated, breaking each chromosome into fragments whose size depends on the radiation dose. These fragments are rescued by fusion with cultured rodent cells, yielding subclones containing different fragments of the human genome. This technique is described by Benham *et al.* (1989) and Cox *et al.*, (1990), which disclosures are hereby incorporated by reference in their entireties. The random and independent nature of the subclones permits efficient mapping of any human genome marker. Human DNA isolated from a panel of 80-

100 cell lines provides a mapping reagent for ordering GENSET cDNAs or genomic DNAs. In this approach, the frequency of breakage between markers is used to measure distance, allowing construction of fine resolution maps as has been done using conventional ESTs (Schuler *et al.*, 1996), which disclosure is hereby incorporated by reference in its entirety.

RH mapping has been used to generate a high-resolution whole genome radiation hybrid map of human chromosome 17q22-q25.3 across the genes for growth hormone (GH) and thymidine kinase (TK) (Foster *et al.*, 1996), the region surrounding the Gorlin syndrome gene (Obermayr *et al.*, 1996), 60 loci covering the entire short arm of chromosome 12 (Raeymaekers *et al.*, 1995), the region of human chromosome 22 containing the neurofibromatosis type 2 locus (Frazer *et al.*, 1992) and 13 loci on the long arm of chromosome 5 (Warrington *et al.*, 1991), which disclosures are hereby incorporated by reference in their entireties.

Mapping of cDNAs to Human Chromosomes using PCR techniques

GENSET cDNAs and genomic DNAs may be assigned to human chromosomes using PCR based methodologies. In such approaches, oligonucleotide primer pairs are designed from the cDNA sequence to minimize the chance of amplifying through an intron. Preferably, the oligonucleotide primers are 18-23 bp in length and are designed for PCR amplification. The creation of PCR primers from known sequences is well known to those with skill in the art. For a review of PCR technology see Erlich (1992), which disclosure is hereby incorporated by reference in its entirety.

The primers are used in polymerase chain reactions (PCR) to amplify templates from total human genomic DNA. PCR conditions are as follows: 60 ng of genomic DNA is used as a template for PCR with 80 ng of each oligonucleotide primer, 0.6 unit of Taq polymerase, and 1 uCu of a ^{32}P-labeled deoxycytidine triphosphate. The PCR is performed in a microplate thermocycler (Techne) under the following conditions: 30 cycles of 94 degree Celsius, 1.4 min; 55 degree Celsius, 2 min; and 72 degree Celsius, 2 min; with a final extension at 72 degree Celsius for 10 min. The amplified products are analyzed on a 6% polyacrylamide sequencing gel and visualized by autoradiography. If the length of the resulting PCR product is identical to the distance between the ends of the primer sequences in the cDNA from which the primers are derived, then the PCR reaction is repeated with DNA templates from two panels of human-rodent somatic cell hybrids, BIOS PCRable DNA (BIOS Corporation) and NIGMS Human-Rodent Somatic Cell Hybrid Mapping Panel Number 1 (NIGMS, Camden, NJ).

PCR is used to screen a series of somatic cell hybrid cell lines containing defined sets of human chromosomes for the presence of a given cDNA or genomic DNA. DNA is isolated from the somatic hybrids and used as starting templates for PCR reactions using the primer pairs from the GENSET cDNAs or genomic DNAs. Only those somatic cell hybrids with chromosomes containing the human gene corresponding to the GENSET cDNA or genomic DNA will yield an

amplified fragment. The GENSET cDNAs or genomic DNAs are assigned to a chromosome by analysis of the segregation pattern of PCR products from the somatic hybrid DNA templates. The single human chromosome present in all cell hybrids that give rise to an amplified fragment is the chromosome containing that GENSET cDNA or genomic DNA. For a review of techniques and analysis of results from somatic cell gene mapping experiments, see Ledbetter *et al.*, (1990), which disclosure is hereby incorporated by reference in its entirety.

Mapping of cDNAs to Chromosomes Using Fluorescence in situ Hybridization

Fluorescence *in situ* hybridization allows the GENSET cDNA or genomic DNA to be mapped to a particular location on a given chromosome. The chromosomes to be used for fluorescence in situ hybridization techniques may be obtained from a variety of sources including cell cultures, tissues, or whole blood.

In a preferred embodiment, chromosomal localization of a GENSET cDNA or genomic DNA is obtained by FISH as described by Cherif *et al.* (1990), which disclosure is hereby incorporated by reference in its entirety. Metaphase chromosomes are prepared from phytohemagglutinin (PHA)-stimulated blood cell donors. PHA-stimulated lymphocytes from healthy males are cultured for 72 h in RPMI-1640 medium. For synchronization, methotrexate (10 uM) is added for 17 h, followed by addition of 5-bromodeoxyuridine (5-BudR, 0.1 mM) for 6 h. Colcemid (1 ug/ml) is added for the last 15 min before harvesting the cells. Cells are collected, washed in RPMI, incubated with a hypotonic solution of KCl (75 mM) at 37 degree Celsius for 15 min and fixed in three changes of methanol:acetic acid (3:1). The cell suspension is dropped onto a glass slide and air dried. The GENSET cDNA or genomic DNA is labeled with biotin-16 dUTP by nick translation according to the manufacturer's instructions (Bethesda Research Laboratories, Bethesda, MD), purified using a Sephadex G-50 column (Pharmacia, Upssala, Sweden) and precipitated. Just prior to hybridization, the DNA pellet is dissolved in hybridization buffer (50% formamide, 2 X SSC, 10% dextran sulfate, 1 mg/ml sonicated salmon sperm DNA, pH 7) and the probe is denatured at 70 degree Celsius for 5-10 min.

Slides kept at -20 degree Celsius are treated for 1 h at 37 degree Celsius with RNase A (100 ug/ml), rinsed three times in 2 X SSC and dehydrated in an ethanol series. Chromosome preparations are denatured in 70% formamide, 2 X SSC for 2 min at 70 degree Celsius, then dehydrated at 4 degree Celsius. The slides are treated with proteinase K (10 ug/100 ml in 20 mM Tris-HCl, 2 mM $CaCl_2$) at 37 degree Celsius for 8 min and dehydrated. The hybridization mixture containing the probe is placed on the slide, covered with a coverslip, sealed with rubber cement and incubated overnight in a humid chamber at 37 degree Celsius. After hybridization and post-hybridization washes, the biotinylated probe is detected by avidin-FITC and amplified with additional layers of biotinylated goat anti-avidin and avidin-FITC. For chromosomal localization, fluorescent R-bands are obtained as previously described (Cherif *et al.*, 1990). The slides are

observed under a LEICA fluorescence microscope (DMRXA). Chromosomes are counterstained with propidium iodide and the fluorescent signal of the probe appears as two symmetrical yellow-green spots on both chromatids of the fluorescent R-band chromosome (red). Thus, a particular GENSET cDNA or genomic DNA may be localized to a particular cytogenetic R-band on a given chromosome.

Use of cDNAs to Construct or Expand Chromosome Maps

Once the GENSET cDNAs or genomic DNAs have been assigned to particular chromosomes using any technique known to those skilled in the art those skilled in the art, particularly those described herein, they may be utilized to construct a high resolution map of the chromosomes on which they are located or to identify the chromosomes in a sample.

Chromosome mapping involves assigning a given unique sequence to a particular chromosome as described above. Once the unique sequence has been mapped to a given chromosome, it is ordered relative to other unique sequences located on the same chromosome. One approach to chromosome mapping utilizes a series of yeast artificial chromosomes (YACs) bearing several thousand long inserts derived from the chromosomes of the organism from which the GENSET cDNAs or genomic DNAs are obtained. This approach is described in Nagaraja *et al.* (1997), which disclosure is hereby incorporated by reference in its entirety. Briefly, in this approach each chromosome is broken into overlapping pieces which are inserted into the YAC vector. The YAC inserts are screened using PCR or other methods to determine whether they include the GENSET cDNA or genomic DNA whose position is to be determined. Once an insert has been found which includes the GENSET cDNA or genomic DNA, the insert can be analyzed by PCR or other methods to determine whether the insert also contains other sequences known to be on the chromosome or in the region from which the GENSET cDNA or genomic DNA was derived. This process can be repeated for each insert in the YAC library to determine the location of each of the GENSET cDNA or genomic DNA relative to one another and to other known chromosomal markers. In this way, a high resolution map of the distribution of numerous unique markers along each of the organisms chromosomes may be obtained.

Identification of genes associated with hereditary diseases or drug response

This example illustrates an approach useful for the association of GENSET cDNAs or genomic DNAs with particular phenotypic characteristics. In this example, a particular GENSET cDNA or genomic DNA is used as a test probe to associate that GENSET cDNA or genomic DNA with a particular phenotypic characteristic.

GENSET cDNAs or genomic DNAs are mapped to a particular location on a human chromosome using techniques such as those described herein or other techniques known in the art. A search of Mendelian Inheritance in Man (V. McKusick, Mendelian Inheritance in Man (available on line through Johns Hopkins University Welch Medical Library) reveals the region of the human

chromosome which contains the GENSET cDNA or genomic DNA to be a very gene rich region containing several known genes and several diseases or phenotypes for which genes have not been identified. The gene corresponding to this GENSET cDNA or genomic DNA thus becomes an immediate candidate for each of these genetic diseases.

Cells from patients with these diseases or phenotypes are isolated and expanded in culture. PCR primers from the GENSET cDNA or genomic DNA are used to screen genomic DNA, mRNA or cDNA obtained from the patients. GENSET cDNAs or genomic DNAs that are not amplified in the patients can be positively associated with a particular disease by further analysis. Alternatively, the PCR analysis may yield fragments of different lengths when the samples are derived from an individual having the phenotype associated with the disease than when the sample is derived from a healthy individual, indicating that the gene containing the cDNA may be responsible for the genetic disease.

Uses of polynucleotides in recombinant vectors

The present invention also relates to recombinant vectors, which include the isolated polynucleotides of the present invention, or fragments thereof and to host cells recombinant for a polynucleotide of the invention, such as the above vectors, as well as to methods of making such vectors and host cells and for using them for production of GENSET polypeptides by recombinant techniques.

Recombinant Vectors

The term "vector" is used herein to designate either a circular or a linear DNA or RNA molecule, which is either double-stranded or single-stranded, and which comprise at least one polynucleotide of interest that is sought to be transferred in a cell host or in a unicellular or multicellular host organism. The present invention encompasses a family of recombinant vectors that comprise a regulatory polynucleotide and/or a coding polynucleotide derived from either the GENSET genomic sequence or the cDNA sequence. Generally, a recombinant vector of the invention may comprise any of the polynucleotides described herein, including regulatory sequences, coding sequences and polynucleotide constructs, as well as any GENSET primer or probe as defined herein.

In a first preferred embodiment, a recombinant vector of the invention is used to amplify the inserted polynucleotide derived from a GENSET genomic sequence or a GENSET cDNA, for example any cDNA selected from the group consisting of sequences of SEQ ID Nos: 1-241, sequences of clone inserts of the deposited clone pool, variants and fragments thereof in a suitable cell host, this polynucleotide being amplified at every time that the recombinant vector replicates.

A second preferred embodiment of the recombinant vectors according to the invention comprises expression vectors comprising either a regulatory polynucleotide or a coding nucleic acid of the invention, or both. Within certain embodiments, expression vectors are employed to express

a GENSET polypeptide which can be then purified and, for example be used in ligand screening assays or as an immunogen in order to raise specific antibodies directed against the GENSET protein. In other embodiments, the expression vectors are used for constructing transgenic animals and also for gene therapy. Expression requires that appropriate signals are provided in the vectors, said signals including various regulatory elements, such as enhancers/promoters from both viral and mammalian sources that drive expression of the genes of interest in host cells. Dominant drug selection markers for establishing permanent, stable cell clones expressing the products are generally included in the expression vectors of the invention, as they are elements that link expression of the drug selection markers to expression of the polypeptide.

More particularly, the present invention relates to expression vectors which include nucleic acids encoding a GENSET protein, preferably a GENSET protein with an amino acid sequence selected from the group consisting of sequences of SEQ ID Nos: 242-482, mature polypeptides included in sequences of SEQ ID Nos: 242-272 and 274-384, and sequences of full-length or mature polypeptides encoded by the clone inserts of the deposited clone pool, as well as variants and fragments thereof. The polynucleotides of the present invention may be used to express an encoded protein in a host organism to produce a beneficial effect. In such procedures, the encoded protein may be transiently expressed in the host organism or stably expressed in the host organism. The encoded protein may have any of the activities described herein. The encoded protein may be a protein which the host organism lacks or, alternatively, the encoded protein may augment the existing levels of the protein in the host organism.

Some of the elements which can be found in the vectors of the present invention are described in further detail in the following sections.

General features of the expression vectors of the invention

A recombinant vector according to the invention comprises, but is not limited to, a YAC (Yeast Artificial Chromosome), a BAC (Bacterial Artificial Chromosome), a phage, a phagemid, a cosmid, a plasmid or even a linear DNA molecule which may comprise a chromosomal, non-chromosomal, semi-synthetic and synthetic DNA. Such a recombinant vector can comprise a transcriptional unit comprising an assembly of:

(1) a genetic element or elements having a regulatory role in gene expression, for example promoters or enhancers. Enhancers are cis-acting elements of DNA, usually from about 10 to 300 bp in length that act on the promoter to increase the transcription.

(2) a structural or coding sequence which is transcribed into mRNA and eventually translated into a polypeptide, said structural or coding sequence being operably linked to the regulatory elements described in (1); and

(3) appropriate transcription initiation and termination sequences. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling

extracellular secretion of translated protein by a host cell. Alternatively, when a recombinant protein is expressed without a leader or transport sequence, it may include a N-terminal residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

Generally, recombinant expression vectors will include origins of replication, selectable markers permitting transformation of the host cell, and a promoter derived from a highly expressed gene to direct transcription of a downstream structural sequence. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably a leader sequence capable of directing secretion of the translated protein into the periplasmic space or the extracellular medium. In a specific embodiment wherein the vector is adapted for transfecting and expressing desired sequences in mammalian host cells, preferred vectors will comprise an origin of replication in the desired host, a suitable promoter and enhancer, and also any necessary ribosome binding sites, polyadenylation signal, splice donor and acceptor sites, transcriptional termination sequences, and 5'-flanking non-transcribed sequences. DNA sequences derived from the SV40 viral genome, for example SV40 origin, early promoter, enhancer, splice and polyadenylation signals may be used to provide the required non-transcribed genetic elements.

The *in vivo* expression of a GENSET polypeptide of the present invention may be useful in order to correct a genetic defect related to the expression of the native gene in a host organism or to the production of a biologically inactive GENSET protein. Consequently, the present invention also comprises recombinant expression vectors mainly designed for the *in vivo* production of a GENSET polypeptide of the present invention by the introduction of the appropriate genetic material in the organism or the patient to be treated. This genetic material may be introduced *in vitro* in a cell that has been previously extracted from the organism, the modified cell being subsequently reintroduced in the said organism, directly *in vivo* into the appropriate tissue.

Regulatory Elements

The suitable promoter regions used in the expression vectors according to the present invention are chosen taking into account the cell host in which the heterologous gene has to be expressed. The particular promoter employed to control the expression of a nucleic acid sequence of interest is not believed to be important, so long as it is capable of directing the expression of the nucleic acid in the targeted cell. Thus, where a human cell is targeted, it is preferable to position the nucleic acid coding region adjacent to and under the control of a promoter that is capable of being expressed in a human cell, such as, for example, a human or a viral promoter.

A suitable promoter may be heterologous with respect to the nucleic acid for which it controls the expression or alternatively can be endogenous to the native polynucleotide containing the coding sequence to be expressed. Additionally, the promoter is generally heterologous with

respect to the recombinant vector sequences within which the construct promoter/coding sequence has been inserted.

Promoter regions can be selected from any desired gene using, for example, CAT (chloramphenicol transferase) vectors and more preferably pKK232-8 and pCM7 vectors.

Preferred bacterial promoters are the LacI, LacZ, the T3 or T7 bacteriophage RNA polymerase promoters, the gpt, lambda PR, PL and trp promoters (EP 0036776), the polyhedrin promoter, or the p10 protein promoter from baculovirus (Kit Novagen), (Smith *et al.*, 1983; O'Reilly *et al.*, 1992), which disclosures are hereby incorporated by reference in their entireties, the lambda PR promoter or also the trc promoter.

Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-L. Selection of a convenient vector and promoter is well within the level of ordinary skill in the art. The choice of a promoter is well within the ability of a person skilled in the field of genetic engineering. For example, one may refer to the book of Sambrook *et al.*, (1989) or also to the procedures described by Fuller *et al.*, (1996), which disclosures are hereby incorporated by reference in their entireties.

Other regulatory elements

Where a cDNA insert is employed, one will typically desire to include a polyadenylation signal to effect proper polyadenylation of the gene transcript. The nature of the polyadenylation signal is not believed to be crucial to the successful practice of the invention, and any such sequence may be employed such as human growth hormone and SV40 polyadenylation signals. Also contemplated as an element of the expression cassette is a terminator. These elements can serve to enhance message levels and to minimize read through from the cassette into other sequences.

Selectable Markers

Selectable markers confer an identifiable change to the cell permitting easy identification of cells containing the expression construct. The selectable marker genes for selection of transformed host cells are preferably dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, TRP1 for *S. cerevisiae* or tetracycline, rifampicin or ampicillin resistance in *E. Coli*, or levan saccharase for mycobacteria, this latter marker being a negative selection marker.

Preferred Vectors

Bacterial vectors

As a representative but non-limiting example, useful expression vectors for bacterial use can comprise a selectable marker and a bacterial origin of replication derived from commercially available plasmids comprising genetic elements of pBR322 (ATCC 37017). Such commercial

vectors include, for example, pKK223-3 (Pharmacia, Uppsala, Sweden), and pGEM1 (Promega Biotec, Madison, WI, USA).

Large numbers of other suitable vectors are known to those of skill in the art, and commercially available, such as the following bacterial vectors: pQE70, pQE60, pQE-9 (Qiagen), pbs, pD10, phagescript, psiX174, pbluescript SK, pbsks, pNH8A, pNH16A, pNH18A, pNH46A (Stratagene); ptrc99a, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia); pWLNEO, pSV2CAT, pOG44, pXT1, pSG (Stratagene); pSVK3, pBPV, pMSG, pSVL (Pharmacia); pQE-30 (QIAexpress).

Bacteriophage vectors

The P1 bacteriophage vector may contain large inserts ranging from about 80 to about 100 kb. The construction of P1 bacteriophage vectors such as p158 or p158/neo8 are notably described by Sternberg (1992, 1994), which disclosure is hereby incorporated by reference in its entirety. Recombinant P1 clones comprising GENSET nucleotide sequences may be designed for inserting large polynucleotides of more than 40 kb (See Linton *et al.*, 1993), which disclosure is hereby incorporated by reference in its entirety. To generate P1 DNA for transgenic experiments, a preferred protocol is the protocol described by McCormick *et al.* (1994), which disclosure is hereby incorporated by reference in its entirety. Briefly, *E. coli* (preferably strain NS3529) harboring the P1 plasmid are grown overnight in a suitable broth medium containing 25 μg/ml of kanamycin. The P1 DNA is prepared from the *E. coli* by alkaline lysis using the Qiagen Plasmid Maxi kit (Qiagen, Chatsworth, CA, USA), according to the manufacturer's instructions. The P1 DNA is purified from the bacterial lysate on two Qiagen-tip 500 columns, using the washing and elution buffers contained in the kit. A phenol/chloroform extraction is then performed before precipitating the DNA with 70% ethanol. After solubilizing the DNA in TE (10 mM Tris-HCl, pH 7.4, 1 mM EDTA), the concentration of the DNA is assessed by spectrophotometry.

When the goal is to express a P1 clone comprising GENSET nucleotide sequences in a transgenic animal, typically in transgenic mice, it is desirable to remove vector sequences from the P1 DNA fragment, for example by cleaving the P1 DNA at rare-cutting sites within the P1 polylinker (*Sfi*I, *Not*I or *Sal*I). The P1 insert is then purified from vector sequences on a pulsed-field agarose gel, using methods similar to those originally reported for the isolation of DNA from YACs (*See e. g.*, Schedl *et al.*, 1993a; Peterson *et al.*, 1993), which disclosures are hereby incorporated by reference in their entireties. At this stage, the resulting purified insert DNA can be concentrated, if necessary, on a Millipore Ultrafree-MC Filter Unit (Millipore, Bedford, MA, USA – 30,000 molecular weight limit) and then dialyzed against microinjection buffer (10 mM Tris-HCl, pH 7.4; 250 μM EDTA) containing 100 mM NaCl, 30 μM spermine, 70 μM spermidine on a microdyalisis membrane (type VS, 0.025 μM from Millipore). The intactness of the purified P1

DNA insert is assessed by electrophoresis on 1% agarose (Sea Kem GTG; FMC Bio-products) pulse-field gel and staining with ethidium bromide.

Viral vectors

In one specific embodiment, the vector is derived from an adenovirus. Preferred adenovirus vectors according to the invention are those described by Feldman and Steg (1996), or Ohno *et al.*, (1994), which disclosures are hereby incorporated by reference in their entireties. Another preferred recombinant adenovirus according to this specific embodiment of the present invention is the human adenovirus type 2 or 5 (Ad 2 or Ad 5) or an adenovirus of animal origin (French patent application No. FR-93.05954), which disclosure is hereby incorporated by reference in its entirety.

Retrovirus vectors and adeno-associated virus vectors are generally understood to be the recombinant gene delivery systems of choice for the transfer of exogenous polynucleotides *in vivo* , particularly to mammals, including humans. These vectors provide efficient delivery of genes into cells, and the transferred nucleic acids are stably integrated into the chromosomal DNA of the host. Particularly preferred retroviruses for the preparation or construction of retroviral *in vitro* or *in vitro* gene delivery vehicles of the present invention include retroviruses selected from the group consisting of Mink-Cell Focus Inducing Virus, Murine Sarcoma Virus, Reticuloendotheliosis virus and Rous Sarcoma virus. Particularly preferred Murine Leukemia Viruses include the 4070A and the 1504A viruses, Abelson (ATCC No VR-999), Friend (ATCC No VR-245), Gross (ATCC No VR-590), Rauscher (ATCC No VR-998) and Moloney Murine Leukemia Virus (ATCC No VR-190; PCT Application No WO 94/24298). Particularly preferred Rous Sarcoma Viruses include Bryan high titer (ATCC Nos VR-334, VR-657, VR-726, VR-659 and VR-728). Other preferred retroviral vectors are those described in Roth *et al.* (1996), PCT Application No WO 93/25234, PCT Application No WO 94/ 06920, Roux *et al.*, (1989), Julan *et al.*, (1992), and Neda *et al.*, (1991), which disclosures are hereby incorporated by reference in their entireties.

Yet another viral vector system that is contemplated by the invention comprises the adeno-associated virus (AAV). The adeno-associated virus is a naturally occurring defective virus that requires another virus, such as an adenovirus or a herpes virus, as a helper virus for efficient replication and a productive life cycle (Muzyczka *et al.*, 1992), which disclosure is hereby incorporated by reference in its entirety. It is also one of the few viruses that may integrate its DNA into non-dividing cells, and exhibits a high frequency of stable integration (Flotte *et al.* 1992; Samulski *et al.*, 1989; McLaughlin *et al.*, 1989), which disclosures are hereby incorporated by reference in their entireties. One advantageous feature of AAV derives from its reduced efficacy for transducing primary cells relative to transformed cells.

BAC vectors

The bacterial artificial chromosome (BAC) cloning system (Shizuya *et al.*, 1992), which disclosure is hereby incorporated by reference in its entirety, has been developed to stably maintain

large fragments of genomic DNA (100-300 kb) in *E. coli*. A preferred BAC vector comprises a pBeloBAC11 vector that has been described by Kim *et al.* (1996), which disclosure is hereby incorporated by reference in its entirety. BAC libraries are prepared with this vector using size-selected genomic DNA that has been partially digested using enzymes that permit ligation into either the *Bam* HI or *Hind*III sites in the vector. Flanking these cloning sites are T7 and SP6 RNA polymerase transcription initiation sites that can be used to generate end probes by either RNA transcription or PCR methods. After the construction of a BAC library in *E. coli*, BAC DNA is purified from the host cell as a supercoiled circle. Converting these circular molecules into a linear form precedes both size determination and introduction of the BACs into recipient cells. The cloning site is flanked by two *Not* I sites, permitting cloned segments to be excised from the vector by *Not* I digestion. Alternatively, the DNA insert contained in the pBeloBAC11 vector may be linearized by treatment of the BAC vector with the commercially available enzyme lambda terminase that leads to the cleavage at the unique *cos*N site, but this cleavage method results in a full length BAC clone containing both the insert DNA and the BAC sequences.

Baculovirus:

Another specific suitable host vector system is the pVL1392/1393 baculovirus transfer vector (Pharmingen) that is used to transfect the SF9 cell line (ATCC No. CRL 1711) which is derived from *Spodoptera frugiperda*. Other suitable vectors for the expression of the GENSET polypeptide of the present invention in a baculovirus expression system include those described by Chai *et al.*, (1993), Vlasak *et al.*, (1983), and Lenhard *et al.*, (1996), which disclosures are hereby incorporated by reference in their entireties.

Delivery Of The Recombinant Vectors:

To effect expression of the polynucleotides and polynucleotide constructs of the invention, these constructs must be delivered into a cell. This delivery may be accomplished *in vitro*, as in laboratory procedures for transforming cell lines, or *in vivo* or *ex vivo*, as in the treatment of certain diseases states. One mechanism is viral infection where the expression construct is encapsulated in an infectious viral particle.

Several non-viral methods for the transfer of polynucleotides into cultured mammalian cells are also contemplated by the present invention, and include, without being limited to, calcium phosphate precipitation (Graham *et al.*, 1973; Chen *et al.*, 1987); DEAE-dextran (Gopal, 1985); electroporation (Tur-Kaspa *et al.*, 1986; Potter *et al.*, 1984); direct microinjection (Harland *et al.*, 1985); DNA-loaded liposomes (Nicolau *et al.*, 1982; Fraley *et al.*, 1979); and receptor-mediated transfection. (Wu and Wu, 1987, 1988), which disclosures are hereby incorporated by reference in their entireties. Some of these techniques may be successfully adapted for *in vivo* or *ex vivo* use.

Once the expression polynucleotide has been delivered into the cell, it may be stably integrated into the genome of the recipient cell. This integration may be in the cognate location and

orientation via homologous recombination (gene replacement) or it may be integrated in a random, non-specific location (gene augmentation). In yet further embodiments, the nucleic acid may be stably maintained in the cell as a separate, episomal segment of DNA. Such nucleic acid segments or "episomes" encode sequences sufficient to permit maintenance and replication independent of or in synchronization with the host cell cycle.

One specific embodiment for a method for delivering a protein or peptide to the interior of a cell of a vertebrate *in vivo* comprises the step of introducing a preparation comprising a physiologically acceptable carrier and a naked polynucleotide operatively coding for the polypeptide of interest into the interstitial space of a tissue comprising the cell, whereby the naked polynucleotide is taken up into the interior of the cell and has a physiological effect. This is particularly applicable for transfer *in vitro* but it may be applied to *in vivo* as well.

Compositions for use *in vitro* and *in vivo* comprising a "naked" polynucleotide are described in PCT application No. WO 90/11092 (Vical Inc.) and also in PCT application No. WO 95/11307 (Institut Pasteur, INSERM, Université d'Ottawa) as well as in the articles of Tacson *et al.* (1996) and of Huygen *et al.*, (1996), which disclosures are hereby incorporated by reference in their entireties.

In still another embodiment of the invention, the transfer of a naked polynucleotide of the invention, including a polynucleotide construct of the invention, into cells may be proceeded with a particle bombardment (biolistic), said particles being DNA-coated microprojectiles accelerated to a high velocity allowing them to pierce cell membranes and enter cells without killing them, such as described by Klein *et al.*, (1987), which disclosure is hereby incorporated by reference in its entirety.

In a further embodiment, the polynucleotide of the invention may be entrapped in a liposome (Ghosh and Bacchawat, 1991; Wong *et al.*, 1980; Nicolau *et al.*, 1987), which disc' are hereby incorporated by reference in their entireties.

In a specific embodiment, the invention provides a composition for the *in vi* the GENSET protein or polypeptide described herein. It comprises a naked pol operatively coding for this polypeptide, in solution in a physiologically acc suitable for introduction into a tissue to cause cells of the tissue to expr polypeptide.

The amount of vector to be injected to the desired hos of injection. As an indicative dose, it will be injected bet animal body, preferably a mammal body, for example

In another embodiment of the vector acc *vitro* in a host cell, preferably in a host cell pre more preferably a somatic cell such as a muscle ce transformed with the vector coding for the desired GE

c skill integra

thereof is reintroduced into the animal body in order to deliver the recombinant protein within the body either locally or systemically.

Secretion vectors

Some of the GENSET cDNAs or genomic DNAs of the invention may also be used to construct secretion vectors capable of directing the secretion of the proteins encoded by genes inserted in the vectors. Such secretion vectors may facilitate the purification or enrichment of the proteins encoded by genes inserted therein by reducing the number of background proteins from which the desired protein must be purified or enriched. Exemplary secretion vectors are described below.

The secretion vectors of the present invention include a promoter capable of directing gene expression in the host cell, tissue, or organism of interest. Such promoters include the Rous Sarcoma Virus promoter, the SV40 promoter, the human cytomegalovirus promoter, and other promoters familiar to those skilled in the art.

A signal sequence from a polynucleotide of the invention, preferably a signal sequences selected from the group of signal sequences of SEQ ID Nos: 1-31 and 33-143 and signal sequences of clone inserts of the deposited clone pool is operably linked to the promoter such that the mRNA transcribed from the promoter will direct the translation of the signal peptide. The host cell, tissue, or organism may be any cell, tissue, or organism which recognizes the signal peptide encoded by the signal sequence in the GENSET cDNA or genomic DNA. Suitable hosts include mammalian cells, tissues or organisms, avian cells, tissues, or organisms, insect cells, tissues or organisms, or yeast.

In addition, the secretion vector contains cloning sites for inserting genes encoding the proteins which are to be secreted. The cloning sites facilitate the cloning of the insert gene in frame with the signal sequence such that a fusion protein in which the signal peptide is fused to the protein encoded by the inserted gene is expressed from the mRNA transcribed from the promoter. The signal peptide directs the extracellular secretion of the fusion protein.

The secretion vector may be DNA or RNA and may integrate into the chromosome of the host, be stably maintained as an extrachromosomal replicon in the host, be an artificial chromosome, or be transiently present in the host. Preferably, the secretion vector is maintained in multiple copies in each host cell. As used herein, multiple copies means at least 2,5, 10, 20, 25, 50 or more than 50 copies per cell. In some embodiments, the multiple copies are maintained extrachromosomally. In other embodiments, the multiple copies result from amplification of a chromosomal sequence.

Many nucleic acid backbones suitable for use as secretion vectors are known to those skilled in the art, including retroviral vectors, SV40 vectors, Bovine Papilloma Virus vectors, yeast integrating plasmids, yeast episomal plasmids, yeast artificial chromosomes, human artificial

chromosomes, P element vectors, baculovirus vectors, or bacterial plasmids capable of being transiently introduced into the host.

The secretion vector may also contain a polyA signal such that the polyA signal is located downstream of the gene inserted into the secretion vector.

After the gene encoding the protein for which secretion is desired is inserted into the secretion vector, the secretion vector is introduced into the host cell, tissue, or organism using calcium phosphate precipitation, DEAE-Dextran, electroporation, liposome-mediated transfection, viral particles or as naked DNA. The protein encoded by the inserted gene is then purified or enriched from the supernatant using conventional techniques such as ammonium sulfate precipitation, immunoprecipitation, immunochromatography, size exclusion chromatography, ion exchange chromatography, and hplc. Alternatively, the secreted protein may be in a sufficiently enriched or pure state in the supernatant or growth media of the host to permit it to be used for its intended purpose without further enrichment.

The signal sequences may also be inserted into vectors designed for gene therapy. In such vectors, the signal sequence is operably linked to a promoter such that mRNA transcribed from the promoter encodes the signal peptide. A cloning site is located downstream of the signal sequence such that a gene encoding a protein whose secretion is desired may readily be inserted into the vector and fused to the signal sequence. The vector is introduced into an appropriate host cell. The protein expressed from the promoter is secreted extracellularly, thereby producing a therapeutic effect.

Cell Hosts

Another object of the invention comprises a host cell that has been transformed or transfected with one of the polynucleotides described herein, and in particular a polynucleotide either comprising a GENSET regulatory polynucleotide or the polynucleotide coding for a GENSET polypeptide. Also included are host cells that are transformed (prokaryotic cells) or that are transfected (eukaryotic cells) with a recombinant vector such as one of those described above. However, the cell hosts of the present invention can comprise any of the polynucleotides of the present invention. In a preferred embodiment, host cells contain a polynucleotide sequence comprising a sequence selected from the group consisting of sequences of SEQ ID Nos: 1-241, sequences of clone inserts of the deposited clone pool, variants and fragments thereof. Preferred host cells used as recipients for the expression vectors of the invention are the following:

a) Prokaryotic host cells: *Escherichia coli* strains (I.E.DH5-α strain), *Bacillus subtilis*, *Salmonella typhimurium*, and strains from species like *Pseudomonas*, *Streptomyces* and *Staphylococcus*.

b) Eukaryotic host cells: HeLa cells (ATCC No.CCL2; No.CCL2.1; No.CCL2.2), Cv 1 cells (ATCC No.CCL70), COS cells (ATCC No.CRL1650; No.CRL1651), Sf-9 cells (ATCC

No.CRL1711), C127 cells (ATCC No. CRL-1804), 3T3 (ATCC No. CRL-6361), CHO (ATCC No. CCL-61), human kidney 293. (ATCC No. 45504; No. CRL-1573) and BHK (ECACC No. 84100501; No. 84111301).

c) Other mammalian host cells.

The present invention also encompasses primary, secondary, and immortalized homologously recombinant host cells of vertebrate origin, preferably mammalian origin and particularly human origin, that have been engineered to: a) insert exogenous (heterologous) polynucleotides into the endogenous chromosomal DNA of a targeted gene, b) delete endogenous chromosomal DNA, and/or c) replace endogenous chromosomal DNA with exogenous polynucleotides. Insertions, deletions, and/or replacements of polynucleotide sequences may be to the coding sequences of the targeted gene and/or to regulatory regions, such as promoter and enhancer sequences, operably associated with the targeted gene.

In addition to encompassing host cells containing the vector constructs discussed herein, the invention also encompasses primary, secondary, and immortalized host cells of vertebrate origin, particularly mammalian origin, that have been engineered to delete or replace endogenous genetic material (e.g., coding sequence), and/or to include genetic material (e.g., heterologous polynucleotide sequences) that is operably associated with the polynucleotides of the invention, and which activates, alters, and/or amplifies endogenous polynucleotides. For example, techniques known in the art may be used to operably associate heterologous control regions (e.g., promoter and/or enhancer) and endogenous polynucleotide sequences via homologous recombination, see, e.g., U.S. Patent No. 5,641,670, issued June 24, 1997; International Publication No. WO 96/29411, published September 26, 1996; International Publication No. WO 94/12650, published August 4, 1994; Koller *et al.*, (1989); and Zijlstra *et al.* (1989) (The disclosures of each of which are incorporated by reference in their entireties).

The present invention further relates to a method of making a homologously recombinant host cell *in vitro* or *in vivo*, wherein the expression of a targeted gene not normally expressed in the cell is altered. Preferably the alteration causes expression of the targeted gene under normal growth conditions or under conditions suitable for producing the polypeptide encoded by the targeted gene. The method comprises the steps of: (a) transfecting the cell *in vitro* or *in vivo* with a polynucleotide construct, said polynucleotide construct comprising; (i) a targeting sequence; (ii) a regulatory sequence and/or a coding sequence; and (iii) an unpaired splice donor site, if necessary, thereby producing a transfected cell; and (b) maintaining the transfected cell *in vitro* or *in vivo* under conditions appropriate for homologous recombination.

The present invention further relates to a method of altering the expression of a targeted gene in a cell *in vitro* or *in vivo* wherein the gene is not normally expressed in the cell, comprising the steps of: (a) transfecting the cell *in vitro* or *in vivo* with a polynucleotide construct, said polynucleotide construct comprising: (i) a targeting sequence; (ii) a regulatory sequence and/or a

coding sequence; and (iii) an unpaired splice donor site, if necessary, thereby producing a transfected cell; and (b) maintaining the transfected cell *in vitro* or *in vivo* under conditions appropriate for homologous recombination, thereby producing a homologously recombinant cell; and (c) maintaining the homologously recombinant cell *in vitro* or *in vivo* under conditions appropriate for expression of the gene.

The present invention further relates to a method of making a polypeptide of the present invention by altering the expression of a targeted endogenous gene in a cell *in vitro* or *in vivo* wherein the gene is not normally expressed in the cell, comprising the steps of: a) transfecting the cell *in vitro* with a polynucleotide construct, said polynucleotide construct comprising: (i) a targeting sequence; (ii) a regulatory sequence and/or a coding sequence; and (iii) an unpaired splice donor site, if necessary, thereby producing a transfected cell; (b) maintaining the transfected cell *in vitro* or *in vivo* under conditions appropriate for homologous recombination, thereby producing a homologously recombinant cell; and c) maintaining the homologously recombinant cell *in vitro* or *in vivo* under conditions appropriate for expression of the gene thereby making the polypeptide.

The present invention further relates to a polynucleotide construct which alters the expression of a targeted gene in a cell type in which the gene is not normally expressed. This occurs when the polynucleotide construct is inserted into the chromosomal DNA of the target cell, wherein said polynucleotide construct comprises: a) a targeting sequence; b) a regulatory sequence and/or coding sequence; and c) an unpaired splice-donor site, if necessary. Further included are a polynucleotide construct, as described above, wherein said polynucleotide construct further comprises a polynucleotide which encodes a polypeptide and is in-frame with the targeted endogenous gene after homologous recombination with chromosomal DNA.

The compositions may be produced, and methods performed, by techniques known in the art, such as those described in U.S. Patent Nos: 6,054,288; 6,048,729; 6,048,724; 6,048,524; 5,994,127; 5,968,502; 5,965,125; 5,869,239; 5,817,789; 5,783,385; 5,733,761; 5,641,670; 5,580,734 ; International Publication Nos:WO96/29411, WO 94/12650; and scientific articles described by Koller *et al.*, (1994). (The disclosures of each of which are incorporated by reference in their entireties).

The GENSET gene expression in mammalian cells, preferably human cells, may be rendered defective, or alternatively may be altered by replacing the endogenous GENSET gene in the genome of an animal cell by a GENSET polynucleotide according to the invention. These genetic alterations may be generated by homologous recombination using previously described specific polynucleotide constructs.

Mammal zygotes, such as murine zygotes may be used as cell hosts. For example, murine zygotes may undergo microinjection with a purified DNA molecule of interest, for example a purified DNA molecule that has previously been adjusted to a concentration ranging from 1 ng/ml – for BAC inserts- to 3 ng/µl –for P1 bacteriophage inserts- in 10 mM Tris-HCl, pH 7.4, 250 µM

EDTA containing 100 mM NaCl, 30 μM spermine, and70 μM spermidine. When the DNA to be microinjected has a large size, polyamines and high salt concentrations can be used in order to avoid mechanical breakage of this DNA, as described by Schedl *et al* (1993b), which disclosure is hereby incorporated by reference in its entirety.

Any one of the polynucleotides of the invention, including the Polynucleotide constructs described herein, may be introduced in an embryonic stem (ES) cell line, preferably a mouse ES cell line. ES cell lines are derived from pluripotent, uncommitted cells of the inner cell mass of pre-implantation blastocysts. Preferred ES cell lines are the following: ES-E14TG2a (ATCC No.CRL-1821), ES-D3 (ATCC No.CRL1934 and No. CRL-11632), YS001 (ATCC No. CRL-11776), 36.5 (ATCC No. CRL-11116). ES cells are maintained in an uncommitted state by culture in the presence of growth-inhibited feeder cells which provide the appropriate signals to preserve this embryonic phenotype and serve as a matrix for ES cell adherence. Preferred feeder cells are primary embryonic fibroblasts that are established from tissue of day 13- day 14 embryos of virtually any mouse strain, that are maintained in culture, such as described by Abbondanzo *et al.* (1993) and are growth-inhibited by irradiation, such as described by Robertson (1987), or by the presence of an inhibitory concentration of LIF, such as described by Pease and Williams (1990), which disclosures are hereby incorporated by reference in their entireties.

The constructs in the host cells can be used in a conventional manner to produce the gene product encoded by the recombinant sequence.

Following transformation of a suitable host and growth of the host to an appropriate cell density, the selected promoter is induced by appropriate means, such as temperature shift or chemical induction, and cells are cultivated for an additional period. Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification. Microbial cells employed in the expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents. Such methods are well known by the skilled artisan.

Transgenic Animals

The terms "transgenic animals" or "host animals" are used herein to designate animals that have their genome genetically and artificially manipulated so as to include one of the nucleic acids according to the invention. Preferred animals are non-human mammals and include those belonging to a genus selected from *Mus* (e.g. mice), *Rattus* (e.g. rats) and *Oryctogalus* (e.g. rabbits) which have their genome artificially and genetically altered by the insertion of a nucleic acid according to the invention. In one embodiment, the invention encompasses non-human host mammals and animals comprising a recombinant vector of the invention or a GENSET gene disrupted by homologous recombination with a knock out vector.

The transgenic animals of the invention all include within a plurality of their cells a cloned recombinant or synthetic DNA sequence, more specifically one of the purified or isolated nucleic acids comprising a GENSET coding sequence, a GENSET regulatory polynucleotide, a polynucleotide construct, or a DNA sequence encoding an antisense polynucleotide such as described in the present specification.

Generally, a transgenic animal according the present invention comprises any of the polynucleotides, the recombinant vectors and the cell hosts described in the present invention. In a first preferred embodiment, these transgenic animals may be good experimental models in order to study the diverse pathologies related to the dysregulation of the expression of a given GENSET gene, in particular the transgenic animals containing within their genome one or several copies of an inserted polynucleotide encoding a native GENSET protein, or alternatively a mutant GENSET protein.

In a second preferred embodiment, these transgenic animals may express a desired polypeptide of interest under the control of the regulatory polynucleotides of the GENSET gene, leading to high yields in the synthesis of this protein of interest, and eventually to tissue specific expression of the protein of interest.

In a third preferred embodiment, these transgenic animals may express a desired polypeptide of interest fused to a GENSET signal peptide sequence, leading to the secretion of the fusion (chimeric) polypeptide.

The design of the transgenic animals of the invention may be made according to the conventional techniques well known from the one skilled in the art. For more details regarding the production of transgenic animals, and specifically transgenic mice, it may be referred to US Patents Nos 4,873,191, issued Oct. 10, 1989; 5,464,764 issued Nov 7, 1995; and 5,789,215, issued Aug 4, 1998; these documents being herein incorporated by reference to disclose methods producing transgenic mice.

Transgenic animals of the present invention are produced by the application of procedures which result in an animal with a genome that has incorporated exogenous genetic material. The procedure involves obtaining the genetic material which encodes either a GENSET coding sequence, a GENSET regulatory polynucleotide or a DNA sequence encoding a GENSET antisense polynucleotide, or a portion thereof, such as described in the present specification. A recombinant polynucleotide of the invention is inserted into an embryonic or ES stem cell line. The insertion is preferably made using electroporation, such as described by Thomas *et al.* (1987), which disclosure is hereby incorporated by reference in its entirety. The cells subjected to electroporation are screened (e.g. by selection via selectable markers, by PCR or by Southern blot analysis) to find positive cells which have integrated the exogenous recombinant polynucleotide into their genome, preferably via an homologous recombination event. An illustrative positive-negative selection

procedure that may be used according to the invention is described by Mansour *et al.* (1988), which disclosure is hereby incorporated by reference in its entirety.

The positive cells are then isolated, cloned and injected into 3.5 days old blastocysts from mice, such as described by Bradley (1987), which disclosure is hereby incorporated by reference in its entirety. The blastocysts are then inserted into a female host animal and allowed to grow to term. Alternatively, the positive ES cells are brought into contact with embryos at the 2.5 days old 8-16 cell stage (morulae) such as described by Wood *et al.* (1993), or by Nagy *et al.* (1993), which disclosures are hereby incorporated by reference in their entireties, the ES cells being internalized to colonize extensively the blastocyst including the cells which will give rise to the germ line.

The offspring of the female host are tested to determine which animals are transgenic e.g. include the inserted exogenous DNA sequence and which ones are wild type.

Thus, the present invention also concerns a transgenic animal containing a nucleic acid, a recombinant expression vector or a recombinant host cell according to the invention.

Recombinant Cell Lines Derived From The Transgenic Animals Of The Invention:

A further object of the invention comprises recombinant host cells obtained from a transgenic animal described herein. In one embodiment the invention encompasses cells derived from non-human host mammals and animals comprising a recombinant vector of the invention or a GENSET gene disrupted by homologous recombination with a knock out vector.

Recombinant cell lines may be established *in vitro* from cells obtained from any tissue of a transgenic animal according to the invention, for example by transfection of primary cell cultures with vectors expressing *onc*-genes such as SV40 large T antigen, as described by Chou (1989), and Shay *et al.* (1991), which disclosures are hereby incorporated by reference in their entireties.

USES OF POLYPEPTIDES OF THE INVENTION

Proteins containing multimerization domains

The invention relates to compositions and methods using proteins of the invention containing a multimerization domains such as a leucine zipper or a helix loop helix domain.

Proteins of the invention containing a leucine zipper domain, are herein referred to as LZP, such as the ones described in this section and those containing a leucine zipper domain as shown on Table VI, or parts thereof, preferably fragments comprising a leucine zipper domain, or derivative thereof to mediate multimerization of proteins of interest.

The leucine zipper consists of a periodic repetition of leucine residues at every seventh, covering a distance spanning eight helical turns. The segments containing these periodic arrays of leucine residues appear to exist in an alpha-helical conformation, and the leucine side chains extending from one alpha-helix interact with those from a similar alpha helix of a second

polypeptide, facilitating dimerization. The structure formed by cooperation of these two regions forms a coiled coil (O'Shea E.K., Rutkowski R., Kim P.S. *Science* 243:538-542.,1989).

Leucine-zippers contribute to targeting of various proteins (eg. glucose transporters, Asano, et al., *J. Biol. Chem.*, 267, 19636-19641 (1992)) and permit dimerization of various cytoplasmic hormone receptors and enzymes (Forman, et al., *Mol Endocrinol*, 3, 1610-1626 (1989)). Leucine zippers are also a common feature of protein transcription factors, where they permit homo- or heterodimerization resulting in tight binding to DNA strands (for reviews, see Abel, et al., *Nature* 341, 24-25 (1989); Jones, et al., *Cell* 61, 9-11 (1990); Lamb, et al., *Trends in Biochemical Sciences* 16, 417-422 (1991)).

Leucine zippers have been shown to be useful tools in several areas of biotechnology, especially in protein engineering, where their ability to mediate homo-dimerization or hetero-dimerization has found several applications. For example, Bosslet et al have described the use of a pair of leucine zipper for in vitro diagnosis, in particular for the immunochemical detection and determination of an analyte in a biological liquid (US patent 5,643,731) / Tso et al have used leucine zippers for producing bispecific antibody heterodimers (US patent 5,932,448) / Methods of preparing soluble oligomeric proteins using leucine zippers have been described by Conrad et al (US patent 5,965,712), Ciardelli et al (US patent 5,837,816), Spriggs et al (WO9410308) / Leucine zipper forming sequences have been used by Pelletier et al in protein fragment complementation assays to detect biomolecular interactions (WO9834120). Because of their usefulness in biotechnology, it is thus highly interesting to isolate new leucine zipper domains.

The multimerization activity of proteins containing leucine zipper domains may be assayed using any of the assays known to those skilled in the art including circular dichroism spectrum and thermal melting analyses as described in US patent 5,942,433. Alternatively, the leucine zipper motif in LZP could be used by those skilled in art as a "bait protein" in a well established yeast double hybridization system to identify its interacting protein partners in vivo from cDNA library derived from different tissues or cell types of a given organism. Alternatively, LZP or part thereof could be used by those skilled in art in mammalian cell transfection experiments. When fused to a suitable peptide tag such as $[His]_6$ tag in a protein expression vector and introduced into culture cells, this expressed fusion protein can be immunoprecipitated with its potential interacting proteins by using anti-tag peptide antibody. This method could be chosen either to identify the associated partner or to confirm the results obtained by other methods such as those just mentioned.

In a preferred embodiment, the invention relates to compositions and methods of using LZP or part thereof for preparing soluble multimeric proteins, which consist in multimers of fusion proteins containing a leucine zipper fused to a protein of interest, using any technique known to those skilled in the art including those described in international patent WO9410308, which disclosure is hereby incorporated by reference in its entirety. In another preferred embodiment, LZP or derivative thereof is used to produce bispecific antibody heterodimers as described in US

patent 5,932,448, which disclosure is hereby incorporated by reference in its entirety. Briefly, leucine zippers capable of forming heterodimers are respectively linked to epitope binding components with different specificities. Bispecific antibodies are formed by pairwise association of the leucine zippers, forming an heterodimer which links two distinct epitope binding components. In still another preferred embodiment, LZP or part thereof or derivative thereof is used for detection and determination of an analyte in a biological liquid as described in US patent 5,643,731, which disclosure is hereby incorporated by reference in its entirety. Briefly, a first leucine zipper is immobilized on a solid support and the second leucine zipper is coupled to a specific binding partner for an analyte in a biological fluid. The two peptides are then brought into contact thereby immobilizing the binding partner on the solid phase. The biological sample is then contacted with the immobilized binding partner and the amount of analyte in the sample bound to the binding partner determined. In still another preferred embodiment, the LZP or part thereof may be used to synthesize novel nucleic acid binding proteins which are able to multimerize with proteins of interest, for example to inhibit and/or control cellular growth using any genetic engineering technique known to those skilled in the art including the ones described in the US patent 5,942,433, which disclosure is hereby incorporated by reference in its entirety .

In another embodiment, the invention relates to compositions and methods using the LZP or part thereof or derivative thereof in protein fragment complementation assays to detect biomolecular interactions in vivo and in vitro as described in international patent WO9834120, which disclosures is hereby incorporated by reference in its entirety. Such assays may be used to study the equilibrium and kinetic aspects of molecular interactions including protein-protein, protein-nucleic acid, protein-carbohydrate and protein-small molecule interactions, for screening cDNA libraries for binding to a target protein with unknown proteins or libraries of small organic molecules for biological activity.

Still, another object of the present invention relates to the use of the LZP or part thereof for identifying new leucine zipper domains using any techniques for detecting protein-protein interaction known to those skilled in the art. Among the traditional methods which may be employed are co-immunoprecipitation, crosslinking and co-purification through gradients or chromatographic columns of cell lysates. Once isolated as a protein interacting with the LZP, such an intracellular protein can be identified (e.g. its amino acid sequence determined) and can, in turn, be used, in conjunction with standard techniques, to identify other proteins with which it interacts. The amino acid sequence thus obtained may be used as a guide for the generation of oligonucleotide mixtures that can be used to screen for gene sequences encoding such intracellular proteins. Screening may be accomplished, for example, by standard hybridization or PCR techniques. Techniques for the generation of oligonucleotide mixtures and the screening are well-known. (See, e.g., Ausubel *et al.*, eds., *Current Protocols in Molecular Biology*, J.Wiley and Sons (New York,

NY 1993) and PR Protocols: A Guide to Methods and Applications, 1990, Innis, M. et al., eds. Academic Press, Inc., New York).

Alternatively, methods may be employed which result in the simultaneous identification of genes which encode the intracellular proteins that can dimerize with the LZP or part thereof using any technique known to those skilled in the art. These methods include, for example, probing cDNA expression libraries, in a manner similar to the well known technique of antibody probing of lambda.gt11 libraries, using as a probe a labeled version of the LZP or part thereof, or fusion protein, e.g., the LZP or part thereof fused to a marker (e.g., an enzyme, fluor, luminescent protein, or dye), or an Ig-Fc domain (for technical details on screening of cDNA expression libraries, see Ausubel *et al, supra*). Alternatively, another method for the detection of protein interaction in vivo, the two-hybrid system, may be used.

Protein of SEQ ID NO:261 (internal designation 116-054-3-0-E6-CS)

The 233 amino acids protein of SEQ ID NO: 261 encoded by the cDNA of SEQ ID NO: 20 displays two leucine zipper sites at positions 142-163 and 170-191.

It is believed that the protein of SEQ ID NO: 261 is able to dimerize either with itself (homo-dimerisation) or with an heterologous protein (hetero-dimerisation) of interest, through the mediation of its leucine zipper domain. Preferred polypeptides of the invention are polypeptides comprising fragments of SEQ ID NO: 261 from position 142-163 and 170-191, and fragments having any of the biological activities described herein.

Protein of SEQ ID NO:263 (internal designation 116-055-2-0-F7-CS)

The protein of SEQ ID NO: 263 encoded by the cDNA of SEQ ID NO: 22 displays a leucine zipper pattern situated near its its NH2 terminal part (position 15 to 36).

It is believed that the protein of SEQ ID NO: 263 is able to dimerize either with itself (homo-dimerisation) or with an heterologous protein (hetero-dimerisation) of interest, through the mediation of its leucine zipper domain. Preferred polypeptides of the invention are polypeptides comprising fragments of SEQ ID NO: 263 from position 15 to 36, and fragments having any of the biological activities described herein..

Protein of SEQ ID NO:245 (internal designation 105-026-1-0-A5-CS)

The protein of SEQ ID NO:245 encoded by the cDNA of SEQ ID NO:4 displays a leucine zipper pattern situated near its COOH terminal part (position 371 to 392).

It is believed that the protein of SEQ ID NO: 245 is able to dimerize either with itself (homo-dimerisation) or with an heterologous protein (hetero-dimerisation) of interest, through the mediation of its leucine zipper domain. Preferred polypeptides of the invention are polypeptides comprising fragments of SEQ ID NO: 245 from position 371 to 392, and fragments having any of the biological activities described herein.

Protein of SEQ ID NO: 257 (internal designation 106-043-4-0-H3-CS)

The 265-amino-acid-long protein of SEQ ID: 257 encoded by the cDNA of SEQ ID NO: 16 exhibits homology to the Homo sapiens hypothetical protein (Genbank accession number AJ278482). These two proteins are probably the result of an alternative splicing.

The protein of SEQ ID NO: 257 displays a leucine zipper pattern situated from position 155 to 176. Thus, it is believed that the protein of SEQ ID NO: 257 is able to dimerize either with itself (homo-dimerisation) or with an heterologous protein (hetero-dimerisation) of interest, through the mediation of its leucine zipper domain. Preferred polypeptides of the invention are polypeptides comprising leucine zipper domains fragments and fragments having any of the biological activities described herein.

Protein of SEQ ID NO: 314 (internal designation 188-41-1-0-B8-CS.cor)

A growing number of proteins have been shown to undergo post-translational modification by fatty acids that are covalently linked to cysteine residues through a thioester bond. Fatty acid modifications contribute to intracellular protein localization by facilitating membrane binding and also by strengthening protein-protein interactions. Cycles of palmitoylation and depalmitoylation have been described for a number of intracellular proteins, but the relevant enzymes that catalyze these processes have yet to be fully characterized and the full significance of these cycles remains to be elucidated.

Palmitoyl-protein thioesterase-1 (PPT1) is a lysosomal hydrolase that removes long-chain fatty acyl groups from modified cysteine residues in proteins. Mutations in PPT1 have been found to cause the infantile form of neuronal ceroid lipofuscinosis (INCL).

Soyombo and Hofmann (J. Biol. Chem. 272: 27456-27463 [1997]) identified cDNAs encoding PPT2. The deduced PPT2 protein contains 302 amino acids, including a 27-amino acid leader peptide, a sequence motif characteristic of many thioesterases and lipases, and 5 potential N-linked glycosylation sites. PPT2 shares 18% amino acid identity with PPT1. Soyombo and Hofmann tentatively localized the human PPT2 gene to 6p21.3. Northern blot analysis detected a predominant 2.0-kb PPT2 transcript in the human tissues examined, with the highest expression in skeletal muscle; variable amounts of 2.8- and 7.0-kb transcripts were also observed.

Cell fractionation studies indicate that PPT2 is present in the lysosomal fraction. Immunoblot analysis of recombinant PPT2 expressed in mammalian cells showed 6 PPT2 proteins ranging in size from 31 to 42 kDa. Treatment that removes asparagine-linked oligosaccharides resulted in a single major protein of 31 kDa and a minor protein of 33 kDa.

Recombinant PPT2, like PPT1, possesses thioesterase activity and localizes to the lysosome. Since PPT2 could not substitute for PPT1 in correcting the metabolic defect in INCL cells and was unable to remove palmitate groups from palmitoylated proteins, it appears that PPT2 possesses a different substrate specificity than PPT1. Another study, however, was able to show,

after expression of the recombinant protein in a baculovirus system and using cell lysate as substrate, that the protein had S-thioesterase activity with a preference for acyl groups palmitic and myristic acid.

The subject invention provides the protein/polypeptide of SEQ ID NO:314, encoded by the cDNA of SEQ ID NO:73. The invention also provides biologically active fragments of SEQ ID NO:314. In one embodiment, the polypeptides of SEQ ID NO:314 are interchanged with the corresponding polypeptide encoded by the human cDNA of clone 188-41-1-0-B8-CS. "Biologically active fragments" are defined as those peptide or polypeptide fragments having at least one of the biological functions of the full length protein (e.g., removal of long-chain fatty acyl groups from modified cysteine residues in proteins). Compositions of the protein/polypeptide of SEQ ID NO:314, or biologically active fragments thereof, are also provided by the subject invention. These compositions may be made according to methods well known in the art.

The invention also provides variants of the protein of SEQ ID NO:314. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence encoded by SEQ ID NO:73. Variants according to the subject invention also have at least one functional or structural characteristic of the protein of SEQ ID NO:314. The invention also provides biologically active fragments of the variant proteins. Compositions of variants, or biologically active fragments thereof, are also provided by the subject invention. These compositions may be made according to methods well known in the art. Unless otherwise indicated, the methods disclosed herein can be practiced utilizing the protein encoded by SEQ ID NO:73, biologically active fragments of SEQ ID NO:314, variants of SEQ ID NO:314, and biologically active fragments of the variants.

Because of the redundancy of the genetic code, a variety of different DNA sequences can encode the amino acid sequence of SEQ ID NO:314. In a preferred embodiment, SEQ ID NO:314 is encoded by clone 188-41-1-0-B8-CS or the cDNA of SEQ ID NO:73. It is well within the skill of a person trained in the art to create these alternative DNA sequences which encode proteins having the same, or essentially the same, amino acid sequence. These variant DNA sequences are, thus, within the scope of the subject invention. As used herein, reference to "essentially the same" sequence refers to sequences that have amino acid substitutions, deletions, additions, or insertions that do not materially affect biological activity. Fragments retaining one or more characteristic biological activity of the protein encoded by clone 188-41-1-0-B8-CS are also included in this definition.

In one aspect of the subject invention, SEQ ID NO:314, and variants thereof, can be used to generate polyclonal or monoclonal antibodies. Both biologically active and immunogenic fragments of SEQ ID NO:314, or variant proteins, can be used to produce antibodies. Polyclonal and/or monoclonal antibodies can be made according to methods well known to the skilled artisan.

Antibodies produced in accordance with the subject invention can be used in a variety of detection assays known to those skilled in the art.

SEQ ID NO:314 can be used as a marker for identification of lysosome dysfunction in individuals. In this aspect of the subject invention, antibodies specific for SEQ ID NO:314, or fragments thereof, are used in routine immunoassays to screen for the presence or absence of SEQ ID NO:314, or fragments thereof, in samples containing lysosomal contents. The presence or absence of the protein of SEQ ID NO:314 can be used to provide an indication of lysosomal function and is, thus, useful for diagnostic/prognostic identification of lysosomal dysfunction.

The subject invention also provides materials and methods for the screening of individual samples for the presence or absence of nucleic acids encoding the protein of SEQ ID NO:314, or variants thereof. In one embodiment, nucleic acids are provided for hybridization assays, known to those skilled in the art, of mRNA or cDNA. The hybridization assays are performed upon nucleic acid samples obtained, or derived from, an individual with suspected lysosomal dysfunction. The hybridization assays screen for the presence or absence of nucleic acids encoding SEQ ID NO:314, or variants thereof. The presence or absence of such nucleic acids can be used as a predictive/prognostic indicator of disease state or lysosome function.

Nucleic acids of the invention can also be used in gene replacement or gene therapy protocols. This aspect of the subject invention nucleic acids encoding SEQ ID NO:314, or biologically active fragments thereof, can be introduced into cells and implanted into an individual with lysosomal disorders. In one embodiment, genetically engineered macrophage can be used for the treatment regimen (see, for example, Eto and Ohashi [2000] J. Inherit. Metabol. Dis. 23:293-298). Alternatively, autologous cells may be obtained from an individual, transformed with nucleic acid *ex vivo*, expanded *ex vivo*, and reintroduced into the individual. Such methods are well known to the skilled artisan.

Protein of SEQ ID NO:280 (internal designation 160-75-4-0-A9-CS):

The protein of SEQ ID NO:280, encoded by the cDNA of SEQ ID NO:39 and expressed in the fetal brain, is a chromosome 12 paralog of C7orf2, a human protein described as a transmembrane receptor located on chromosome 7 (Heus, H. C., A. Hing, et al. (1999) Genomics **57**(3): 342-51). In addition, this protein is an ortholog of the murine gene LMBR1L, found to be involved in polydactily in mice (Clark, R. M., P. C. Marker, et al. (2000) Genomics 67(1): 19-27). A high level of homology was also found with a gene identified in Fugu rubripes (AF056116), as well as with C. Elegans R05D3.2 (Gellner, K. and S. Brenner (1999) Genome Res 9(3): 251-8).

The 362-amino-acid-long protein of SEQ ID NO:280, encoded by the cDNA of SEQ ID NO:39 is a splice variant of Z64989, located on chromosome 12. The chromosome 12 gene has 6 known variants described in entries AK001356 and AK001651 in genbank and entries A26354, A26375, X27360 and Z64989 in geneseqn. The closest sequence is Z64989, either at the nucleotide

or the protein level. Z64989 is split into 17 exons, of which the protein of the invention contains the last 14. The transcription start site of the cDNA of SEQ ID NO:39 lies within the third intron of Z64989, and the protein of the invention starts at position 128 of Z64989. In addition, 2 potential leucine zippers are present in the protein of the invention (positions 136-157 and 272-293).

Preaxial polydactyly is a congenital hand malformation that includes duplicated thumbs, various forms of triphalangeal thumbs, and duplications of the index finger. Clark et al. (supra) demonstrated the correspondence between the spatial and temporal changes in Lmbr1 expression and the embryonic onset of polydactyly mutant phenotype, suggesting that a downregulation of Lmbr1 results in polydactily. It is likely that the Lmbr1 gene is involved in the patterning of limbs during mammalian development, for example by receiving and transducing a locally secreted ligand in the developing limb.

It is believed that the protein of SEQ ID NO:280 is a paralog of human C7orf2, and is thus a membrane bound protein implicated in the patterning of the mammalian body plan during early development. For example, the protein of the invention may be involved in organizing limb development, as well as in the development of the fetal brain. As such, the activity of the present protein likely influences various cellular processes, including gene expression, cellular growth and proliferation, as well as cellular differentiation. In addition, leucine zippers within the present protein render the protein capable of undergoing specific protein-protein interactions with other leucine-zipper containing proteins, including with itself (i.e. homodimerization). Preferred polypeptides of the invention are fragments of SEQ ID NO:280 having any of the biological activities described herein.

In one embodiment of the present invention, the present protein can be used to identify cells of the fetal brain. For example, the protein of the invention or part thereof may be used to synthesize specific antibodies using any technique known to those skilled in the art. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, such as in forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, etc., or to differentiate different tissue types in a tissue cross-section using immunochemistry. In addition, labeled reagents that can specifically bind to the protein of the invention can be used to visualize cell membranes and the components of the secretory pathway in cells, e.g. the ER and Golgi.

In another embodiment of the present invention, the present protein can be used to diagnose developmental abnormalities, or the potential for such abnormalities, e.g. in a fetus or in adults to determine (i.e. to determine if they are a carrier of a mutant copy of the gene). Individuals found to carry one or two mutant copies of the present gene would be candidates for, e.g. gene therapy or other strategies to correct or compensate for the gene deficiency, or for strategies to ensure that their children would not be carriers of the mutated gene. The characterization of mutations in genes encoding the present protein would also be of great value in understanding the

nature of polydactyly and other developmental disorders, thereby facilitating the development of other strategies for treating and preventing these disorders.

In another embodiment, the present protein is used to modulate gene expression, cell growth and proliferation, and/or cell differentiation in cells in vitro or in vivo. For example, any of these behaviors can be increased or inhibited in cells grown in vitro, e.g. for protein production or for ex vivo therapeutic strategies. In addition, any disease associated with an increase or decrease in any of these cellular behaviors in vivo can be treated or prevented by enhancing or inhibiting the expression or activity of the protein of the invention in cells in vivo.

Proteins of SEQ ID NOs: 309 and 304 (internal designations 188-11-1-0-B3-CS and 187-34-0-0-112-CS)

The proteins of SEQ ID NOs: 309 and 304 are encoded by the cDNAs of SEQ ID NOs: 68 and 63. Accordingly, it will be appreciated that all characteristics and uses of the polypeptides of SEQ ID NOs: 309 and 304 described throughout the present application also pertain to the polypeptides encoded by human cDNA of clones 188-11-1-0-B3-CS and 187-34-0-0-112-CS. In addition, it will be appreciated that all characteristics and uses of the nucleic acids of SEQ ID NOs: 68 and 63 described throughout the present application also pertain to the nucleic acids of the human cDNAs of clones 188-11-1-0-B3-CS and 187-34-0-0-112-CS.

The protein of SEQ ID NO: 309 (encoded by the clone having internal designation number 188-11-1-0-B3-CS) and the polymorphic variant thereof of SEQ ID NO: 304 (encoded by the clone having internal identification number 187-34-0-0-112-CS and which differs from the polypeptide encoded by the clone having internal designation number 188-11-1-0-B3CS at a single amino acid), are highly homologous to the first 279 amino acids of the LGI1 (Leucine-rich gene – Glioma Inactivated) protein. Clones 188-11-1-0-B3-CS and 187-34-0-0-112-CS appear to be splicing and polymorphic variants of LGI1. The LGI1 protein is 557 amino acid in length. (See Somerville et al., (2000) Mammalian Genome 11, 622-627 ; Chernova, et al. (1998) Oncogene 17, 2873-2881, the disclosures of which are incorporated herein by reference in their entireties). Clone 188-11-1-0-B3-CS align with the first 279 amino acids of LGI1, followed by the addition of 12 amino acids (VLREIHRFTNMS) to the C-terminal end which do not appear to be homologous to LGI1. Like LGI1, clone 188-11-1-0-B3-CS and the polymorphic variant 187-34-0-0-112-CS contain the LRR domain and are highly expressed in brain tissue.

LGI1 belongs to a large family of leucine-rich repeat (LRR) proteins. It is believed that the LRR domains act as a region of protein-protein interaction. This has been substantiated as the family of known LRR proteins has grown. Leucine–rich repeats have been identified as essential components in glycoprotein hormone receptors, proteoglycans and the Trk proteins by expression of mutants and artificial chimaeras in tissue culture and by biochemical analysis of the properties of these constructs. Many transmembrane LRR proteins are known or suspected to encode truncated

forms (N and L^6, and slit for example) with functional significance. The proteoglycan Decorin, a secreted protein, binds TGF-β, a growth factor which stimulates decorin expression. Since decorin inhibits growth of cultured cells, it may form part of a negative feedback loop to regulate cell growth. This is similar to the proposed function of the LGI1 receptor protein.

Analysis of brain gliomas has revealed that LGI1 expression is either abolished or greatly reduced in high-grade tumors compared with more benign ones, indicating a role as a tumor suppressor gene (Cowell et al. 2000; Cowell et al. 1998, the disclosure of which is incorporated herein by reference in its entirety). Most glioblastoma multiforme (GBM) brain tumors contain only one genomic copy of LGI1, and this one is almost invariably not expressed. How the gene is inactivated is not clear, although one possibility is that chromosome or gene rearrangement, which occur in 20-25% of tumors, cause inactivation as a result of a positional effect. Recently it was determined that the LGI1 gene is located on 10q24, and is disrupted by translocation in the T98G GBM cell line and is also rearranged in over 26% of primary brain tumors. Alternatively, LGI1 may be part of a highly regulated pathway where inactivation of other key members or high specific transcription factors results in either inactivation of all genes in the pathway or a failure to initiate transcription.

Since functional inactivation of LGI1 occurs during the transition of low-grade to high-grade brain tumors, knockout or transgenic mice in which the expression of the protein of SEQ ID NO:309 or 304 has been reduced, eliminated or altered may be used as disease model. In particular, mice that overexpress LGI1 may be used as a tumorigenesis model.

Mice are particularly useful as models for assessing the consequences of altering the level or activity of the proteins of SEQ ID NO:309 or 304 or to identify agents useful in treaating tumorigenesis, since human and mouse LGI1 are highly conserved, showing 91% identity at the nucleotide level and 97% similarity at the amino acid level, with most of the amino acid substitutions being conservative. The mouse lgi1 gene is 4.2 kb in length, while the human LGI1 is 2.2 kb in length. This difference in size between the human and mouse gene is a result of the inclusion of a 2 kb sequence in the 3' untranslated region in the mouse gene. Whether the additional sequence affects gene expression is not clear. Further analysis of the genomic sequence reveals that the number of exon/intron boundaries is also similar in humans and mice. The high degree of LGI1 conservation between mice and humans implies that this gene has experienced a strong selection pressure. It is intriguing to speculate that any major deviations in the primary protein sequence may result in a loss of function of this gene product. Total or partial loss of the LGI1 gene function could, therefore, be lethal, which in turn implies that LGI1 plays an important role in normal brain development as well as in tumor formation.

SEQ ID NOs:309 and 304 also have high homology with Slit, a secreted Drosophila protein which plays a role in the development of axon pathway development in the central nervous system. The Slit protein is necessary for the normal development of the midline on the CNS, particularly the

midline glial cells, and for the concomitant formation of the commissural axon pathway. The process is dependent on the level of Slit protein expression. It appears that the Slit protein is excreted by the midline glial cells, where it is synthesized and is eventually associated with the surface axons that traverse them. Contact of cells with supernatant expressing the product of this gene increases the permeability of THP-1 monocyte cells to calcium. Thus, it is likely that Slit is involved in a signal transduction pathway that is initiated when Slit protein binds a receptor on the surface of the monocyte cell.

In view of the above, it is believed that the proteins of SEQ ID NOs:309 and 304 are involved in a signal transduction pathway mediated through a receptor that modulates the differentiation and/or proliferation of cells.

Northern blot analysis detects LGI1 transcripts only in brain, neural tissue, and skeletal muscle but not in heart, kidney, lung, placenta, liver, or pancreas. Northern blot analysis of RNA derived from several different regions of human brain revealed a widespread expression of LGI1 although with different intensities. The highest abundance was found in cerebral cortex, hippocampus, and putamen. The lowest expression was detected in corpus callosum. The levels of expression were intermediate in the other brain regions. Accordingly, the proteins of SEQ ID NOs:309 or 304 or fragments thereof, as well as polynucleotides encoding the proteins of SEQ ID NOs:309 or 304, may be used to determine whether a tissue sample is derived from brain (and in particular cerebral cortex, hippocampus, or putamen), neural tissue, and skeletal tissue or to distinguish whether a tissue sample is derived from brain or another tissue, such as heart, kidney, lung, placenta, liver, or pancreas.

Accordingly, the present invention includes the use of the protein of SEQ ID NOs: 309 or 304, fragments comprising at least 5, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, or 200 consecutive amino acids thereof, or fragments having a desired biological activity to treat or ameliorate a condition, such as those listed above, in an individual. In such embodiments, the protein of SEQ ID NO:309 or 304, or a fragment thereof, is administered to an individual in whom it is desired to increase or decrease any of the activities of the protein of SEQ ID NO:309 or 304, including tumor suppression, modulation of neural development or involvement in brain tumors, glioblastoma multiforme, brain injuries, neurodegenerative disease states and behavioral disorders such as Alzheimers Disease, Parkinsons Disease, epilepsy, multiple sclerosis, Huntingtons Disease, schizophrenia, obsessive compulsive disorders, and in the processes of nerve regeneration in spinal cord injury, stroke, facial nerve damage, diabetes caused nerve damage, and retinal regeneration.

The protein of SEQ ID NO:309 or 304 or a fragment thereof may be administered directly to the individual or, alternatively, a nucleic acid encoding the protein of SEQ NO:309 or 304 or a fragment thereof may be administered to the individual. Alternatively, an agent which increases the activity of the protein of SEQ ID NO:309 or 304 may be administered to the individual. Such agents may be identified by contacting the protein of SEQ NO:309 or 304 or a cell or preparation

containing the protein of SEQ ID NO:309 or 304 with a test agent and assaying whether the test agent increases the activity of the protein. For example, the test agent may be a chemical compound or a polypeptide or peptide.

Alternatively, the activity of the protein of SEQ ID NO:309 or 304 may be decreased by administering an agent which interferes with such activity to an individual. Agents which interfere with the activity of the protein of SEQ ID NO:309 or 304 may be identified by contacting the protein or a cell or preparation containing the with a test agent and assaying whether the test agent decreases the activity of the protein. For example, the agent may be a chemical compound, a polypeptide or peptide, an antibody, or a nucleic acid such as an antisense nucleic acid or a triple helix-forming nucleic acid.

In one embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably brain, or to distinguish between two or more possible sources of a tissue sample on the basis of the level of the protein of SEQ ID NO:309 or 304 in the sample. For example, the protein of SEQ ID NO:309 or 304 or fragments thereof may be used to generate antibodies using any techniques known to those skilled in the art, including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples,differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry. In such methods a tissue sample is contacted with the antibody, which may be detectably labeled, under conditions which facilitate antibody binding. The level of antibody binding to the test sample is measured and compared to the level of binding to control cells from brain or tissues other than brain to determine whether the test sample is from brain. Alternatively, the level of the protein of SEQ ID NO:309 or 304 in a test sample may be measured by determining the level of RNA encoding the protein of SEQ ID NO:309 or 304 in the test sample. RNA levels may be measured using nucleic acid arrays or using techniques such as in situ hybridization, Northern blots, dot blots or other technques familiar to those skilled in the art. If desired, an amplification reaction, such as a PCR reaction, may be performed on the nucleic acid sample prior to analysis. The level of RNA in the test sample is compared to RNA levels in control cells from brain or tissues other than brain to determine whether the test sample is from brain. For a number of disorders listed above, particularly of the nervous system, expression of the genes encoding the polyepeptide of SEQ ID NO:309 or 304 at significant higher or lower levels may be routinely detected in certain tissues or cell types (e.g., cancerous and wounded tissues) or bodily fluids (e.g., serum, plasma, synovial fluid, and spinal fluid) or another tissue of cell sample taken from an individual having such a disorder, relative to the standard gene expression level, i.e., the expression level in healthy tissue or bodily fluid from an individual not having the disorder.

In another embodiment, antibodies to the protein of SEQ ID NO:309 or 304 or part thereof may be used for detection, enrichment, or purification of cells expressing the protein of SEQ ID

NO:309 or 304, including using methods known to those skilled in the art. For example, an antibody against the protein of SEQ ID NO:309 or 304 or a fragment thereof may be fixed to a solid support, such as a chromatograpy matrix. A preparation containing cells expressing the protein of SEQ ID NO:309 or 304 is placed in contact with the antibody under conditions which facilitate binding to the antibody. The support is washed and then the cells are released from the support by contacting the support with agents which cause the cells to dissociate from the antibody.

In another embodiment of the present invention, the protein of SEQ ID NO:309 or 304 or a fragment thereof may be used to diagnose disorders associated with altered expression of the protein of SEQ ID NO:309 or 304. In some embodiments, the protein of SEQ ID NO:309 or 304 or fragments thereof may be used to diagnose cancer. In such techniques, the level of the protein of SEQ ID NO:309 or 304 in an ill individual is measured using techniques such as those described herein and compared to the level in normal individuals. For example, a decreased level of the protein of SEQ ID NO:309 or 304 relative to normal individuals suggests that the ill individual may suffer from cancer or be predisposed to getting cancer in the future.

Another embodiment of the present invention is a polypeptide comprising a structural or functional domain of the protein of SEQ ID NO:309 or 304. Such structural or functional domains of the protein of SEQ ID NO:309 or 304 include a leucine rich repeat C-terminal domain located between amino acid positions 173 and 222, a leucine rich repeat located between amino acid positions 92 and 115, a leucine rich repeat located between amino acid positions 116 and 139, a leucine rich repeat located between amino acid positions 140 and 163, a leucine rich repeat located between amino acid positions 164 and 185, a membrane spanning segment located between amino acid positions 15 and 35, and a signal peptide comprising the sequence FLCLLSALLLTEG/KK.

Accordingly, the protein of SEQ ID NO:309 or 304 or fragments thereof, or polynucleotides encoding these proteins or fragments, may be used in *in vitro* diagnostic assays for malignant brain tumors, such as glioblastoma muliforme. These proteins or nucleic acids may also be used in the attenuation / prevention and/or treatment of brain tumors and/or brain injuries, of neurodegenerative disease states and behavioral disorders such as Alzheimers Disease, Parkinsons Disease, epilepsy, multiple sclerosis, Huntingtons Disease, schizophrenia, obsessive compulsive disorders, and in the processes of nerve regeneration in spinal cord injury, stroke, facial nerve damage, diabetes caused nerve damage, and retinal regeneration.

In addition, the protein, as well as, antibodies directed against the protein, and relevant small molecules may be used as tumor markers and /or immunotherapy targets for the above disease states. For example, antibodies directed against amino acids VLREIHRFTNMS of both clones may aid in the differential detection of the secreted and receptor forms of this protein, since the proteins of SEQ ID NOs:309 and 304 have homology to the secreted forms of LGI1. In addition, the proteins of SEQ ID NOs:309 and 304 or fragments thereof may be used to identify binding partners as described herein.

DNA-binding proteins

The invention relates to compositions and methods using proteins of the invention containing a DNA-binding domain, herein referred to as DBP, such as the ones described in this section and those containing a DNA binding domain domain as shown on Table VI, or parts thereof, preferably fragments comprising a DNA binding domain, or derivative thereof.

Transcriptional regulation is primarily achieved by the sequence-specific binding of proteins to DNA and RNA. Of the known protein motifs involved in the sequence specific recognition of DNA, the zinc finger protein is unique in its modular nature. Zinc finger domains are found in numerous zinc binding proteins which are involved in protein-nucleic acid interactions. They are independently folded zinc-containing mini-domains which are used in a modular repeating fashion to achieve sequence-specific recognition of DNA (Klug 1993 Gene 135, 83-92). Such zinc binding proteins are commonly involved in the regulation of gene expression, and usually serve as transcription factors (see US patents 5,866,325; 6,013,453 and 5,861,495).

To date, zinc finger proteins have been identified which contain between 2 and 37 modules. More than two hundred proteins, many of them transcription factors, have been shown to possess zinc fingers domains. Zinc fingers connect transcription factors to their target genes mainly by binding to specific sequences of DNA. Zinc finger modules are found in a wide variety of transcription regulatory proteins in eukaryotic organisms. A zinc finger domain is generally composed of 25 to 30 amino acid residues which form one or more tetrahedral ion binding sites. The binding sites contain four ligands consisting of the sidechains of cysteine, histidine and occasionally aspartate or glutamate. The binding of zinc allows the relatively short stretches of polypeptide to fold into defined structural units which are well-suited to participate in macromolecular interactions (Berg, J. M. et al. (1996) Science 271:1081-1085). The zinc finger domain was first recognized in the transcription factor TFfIIIA from Xenopus oocytes (Miller, et al., EMBO, 4:1609-1614, 1985; Brown, et al., FEBS Lett., 186:271-274, (1985)).

Zinc binding domains which contain a C_3HC_4 sequence motif are known as RING domains (Lovering, R. et al. (1993) Proc. Natl. Acad. Sci. USA 90:2112-2116). The RING domain consists of eight metal binding residues, and the sequences that bind the two metal ions overlap (Barlow, P. N. et al. (1994) J. Mol. Biol. 237:201-211). Functions of RING finger proteins are mediated through DNA binding and include the regulation of gene expression, DNA recombination, and DNA repair (see Borden and Freemont, Curr Opin Struct Biol 6:395-401 (1996) and US patent 5,861,495).

Both the RING finger and the LIM domain mediate protein-protein interactions and are involved in transcriptional control, either by directly affecting transcription or recruiting co-activators or co-repressors. LIM domains also contribute to various signalling pathways. They may interact with protein kinases and anchor gene products to large protein complexes or to cellular compartments.

PHD fingers are C_4HC_3 zinc fingers spanning approximately 50-80 residues and distinct from RING fingers or LIM domains. They are thought to be mostly DNA or RNA binding domain but may also be involved in protein-protein interactions (for a review see Aasland et al, Trends Biochem Sci

20:56-59 (1995)). The PHD finger domain, belonging to zinc finger domain family, is found in many regulatory proteins which are frequently associated with chromatin-mediated transcriptional regulation.

The nucleic acid binding activity of DBP or part thereof may be assayed using any of the assays known to those skilled in the art including those described in US patent 6,013,453.

The invention relates to compositions and methods using DBPs or part thereof, especially fragments comprising a DNA-binding domain, to stimulate gene transcription.

One of the remarkable features of activation domains of transcriptional factors in general is that "fusing" them to heterologous protein domains seldom affects their ability to activate transcription when recruited to a wide variety of promoters. The high degree of functional independence exhibited by these activation domains makes them valuable tools in various biological assays for analyzing gene expression and protein--protein or protein-RNA or protein-small molecule drug interactions. Several strategies to improve the potency of activation domains and thereby the expression of genes under their control have been reported. These approaches generally involve increasing the number of copies of activation domains fused to the DNA binding domain or generating activators containing synergizing combinations of activation domains.

Therefore, in an additional embodiment, this invention provides compositions and methods containing new transcription factors comprising DBP or part thereof, preferably fragments containing DNA-binding domains. Such transcription factors may be designed to regulate the expression of target genes of interest. Aspects of the invention are applicable to systems involving either covalent or non-covalent linking of the transcription activation domain to a DNA binding domain. In practice, cells can be engineered by the introduction of recombinant nucleic acids encoding the fusion proteins containing at least two mutually heterologous domains, one of them being the DNA-binding domain of the invention, and in some cases additional nucleic acid constructs, to render them capable of ligand-dependent regulation of transcription of a target gene. Administration of the ligand to the cells then regulates (positively, or in some cases, negatively) target gene transcription (all laboratory methods related to this embodiment are completely described in US patents 6.015.709, which disclosure is hereby incorporated by reference in its entirety). Illustrative (non-limiting) example of heterologous domains which can be included along with a DNA-binding domain in various fusion proteins of this invention include another transcription regulatory domains (i.e., transcription activation domains such as a p65, VP16 or AP domain; transcription potentiating or synergizing domains; or transcription repression domains such as an ssn-6/TUP-1 domain or Kruppel family suppressor domain); a DNA binding domain such as a GAL4, lex A or a composite DNA binding domain such as a composite zinc finger domain or a ZFHD1 domain; or a ligand-binding domain comprising or derived from (a) an immunophilin, cyclophilin or FRB domain; (b) an antibiotic binding domain such as tetR: or (c) a hormone receptor such as a progesterone receptor or ecdysone receptor. A wide variety of ligand binding domains may be used in this invention, although ligand binding domains which bind to a cell

permeant ligand are preferred. It is also preferred that the ligand have a molecular weight under about 5 kD, more preferably below 2.5 kD and optimally below about 1500 D. Non-proteinaceous ligands are also preferred. Examples of ligand binding domain/ligand pairs that may be used in the practice of this invention include, but are not limited to: FKBP:FK1012, FKBP:synthetic divalent FKBP ligands (see WO 96/0609 and WO 97/31898), FRB:rapamycin/FKBP (see e.g., WO 96/41865 and Rivera et al, "A humanized system for pharmacologic control of gene expression", Nature Medicine 2(9):1028-1032 (1997)), cyclophilin:cyclosporin (see e.g. WO 94/18317), DHFR:methotrexate (see e.g. Licitra et al, 1996, Proc. Natl. Acad. Sci. U.S.A. 93:12817-12821), TetR:tetracycline or doxycycline or other analogs or mimics thereof (Gossen and Bujard, 1992, Proc. Natl. Acad. Sci. U.S.A. 89:5547; Gossen et al, 1995, Science 268:1766-1769; Kistner et al, 1996, Proc. Natl. Acad. Sci. U.S.A. 93:10933-10938), a progesterone receptor:RU486 (Wang et al, 1994, Proc. Natl. Acad. Sci. U.S.A. 91:8180-8184), ecodysone receptor:ecdysone or muristerone A or other analogs or mimics thereof (No et al, 1996, Proc. Natl. Acad. Sci. U.S.A. 93:3346-3351) and DNA gyrase:coumermycin (see e.g. Farrar et al, 1996, Nature 383:178-181). In many applications it is preferable to use a DNA binding domain which is heterologous to the cells to be engineered. In the case of composite DNA binding domains, component peptide portions which are endogenous to the cells or organism to be engineered are generally preferred.

In another aspect of this embodiment, polynucleotides encoding DNA-binding domains as well as any other functional fragments of DBP may be introduced into polynucleotides encoding fusion proteins for a variety of regulated gene expression systems, including both allostery-based systems such as those regulated by tetracycline, RU486 or ecdysone, or analogs or mimics thereof, and dimerization-based systems such as those regulated by divalent compounds like FK1012, FKCsA, rapamycin, AP1510 or coumermycin, or analogs or mimics thereof, all as described below (See also, Clackson, Controlling mammalian gene expression with small molecules, Current Opinion in Chem. Biol. 1:210-218 (1997)). The fusion proteins may comprise any combination of relevant components, including bundling domains, DNA binding domains, transcription activation (or repression) domains and ligand binding domains. Other heterologous domains may also be included.

Another embodiment of this invention relates to expression systems, preferably vectors and vector-containing cells, using DBP or part thereof, especially the DNA-binding domain. In this regard, recombinant nucleic acids are provided which encode fusion proteins containing the transcription activation domain of the invention and at least one additional domain that is heterologous thereto, where the peptide sequence of said activation domain is itself eventually modified relative to the naturally occurring sequence from which it was derived to increase or decrease its potency as a transcriptional activator relative to the counterpart comprising the native peptide sequence. Each of the recombinant nucleic acids of this invention may further comprise an expression control sequence operably linked to the coding sequence and may be provided within a DNA vector, e.g., for use in transducing prokaryotic or eukaryotic cells. Some of the recombinant nucleic acids of a given

composition as described above, including any optional recombinant nucleic acids, may be present within a single vector or may be apportioned between two or more vectors. The recombinant nucleic acids may be provided as inserts within one or more recombinant viruses which may be used, for example, to transduce cells in vitro or cells present within an organism, including a human or non-human mammalian subject. It should be appreciated that non-viral approaches (naked DNA, liposomes or other lipid compositions, etc.) may be used to deliver recombinant nucleic acids of this invention to cells in a recipient organism. The resultant engineered cells and their progeny containing one or more of these recombinant nucleic acids or nucleic acid compositions of this invention may be used in a variety of important applications, including human gene therapy, analogous veterinary applications, the creation of cellular or animal models (including transgenic applications) and assay applications. Such cells are useful, for example, in methods involving the addition of a ligand, preferably a cell permeant ligand, to the cells (or administration of the ligand to an organism containing the cells) to regulate expression of a target gene.

The invention also relates to methods and compositions using DBP or part thereof to bind to nucleic acids, preferably DNA, alone or in combination with other substances. For example, DBP or part thereof is added to a sample containing nucleic acid in conditions allowing binding, and allowed to bind to nucleic acids. In a preferred embodiment, DBP or part thereof may be used to purify nucleic acids such as restriction fragments. In another preferred embodiment, DBP or part thereof may be used to visualize nucleic acids when the polypeptide is linked to an appropriate fusion partner, or is detected by probing with an antibody. Alternatively, DBP or part thereof may be bound to a chromatographic support, either alone or in combination with other DNA binding proteins, using techniques well known in the art, to form an affinity chromatography column. A sample containing nucleic acids to purify is run through the column. Immobilizing DBP or part thereof on a support advantageous is particularly for those embodiments in which the method is to be practiced on a commercial scale. This immobilization facilitates the removal of the protein from the batch of product and subsequent reuse of the protein. Immobilization of DBP or part thereof can be accomplished, for example, by inserting a cellulose-binding domain in the protein. One of skill in the art will understand that other methods of immobilization could also be used and are described in the available literature.

In another embodiment, the present invention relates to compositions and methods using DBP or part thereof, especially the DNA-binding domain, to alter the expression of genes of interest in a target cells. Such genes of interest may be disease related genes, such as oncogenes or exogenous genes from pathogens, such as bacteria or viruses using any techniques known to those skilled in the art including those described in US patents 5,861,495; 5,866,325 and 6,013,453.

In still another embodiment, DBP or part thereof may be used to diagnose, treat and/or prevent disorders linked to dysregulation of gene transcription such as cancer and other disorders relating to abnormal cellular differentiation, proliferation, or degeneration, including

hyperaldosteronism, hypocortisolism (Addison's disease), hyperthyroidism (Grave's disease), hypothyroidism, colorectal polyps, gastritis, gastric and duodenal ulcers, ulcerative colitis, and Crohn's disease.

Protein of SEQ ID NO: 388 (internal designation 109-002-4-0-C6-CS)

The protein of SEQ ID NO: 388 encoded by cDNA of SEQ ID NO: 147 is a 375 amino-acids long protein containing a zinc finger domain, namely a PHD-finger domain from positions 329 to 339.

The PHD finger was originally identified by comparison of the maize homeodomain (HD) protein ZMHOX1a (Bellmann R. and Werr W. EMBO J. 11: 3367-3374 (1992)) to its Arabidopsis relative HAT3.1 and named plant homeodomain (PHD) finger due to its association with the DNA-binding HD in both genes. This motif often occurs in various regulatory genes, such as members of the trithorax (TRX-G) or polycomb (PC-G) groups (Aasland R. et al. Trends Biochem.Sci. 20: 56-59 (1995)) and leukaemia-associated proteins (LAP finger) (Saha V. et al. Proc.Natl.Acad.Sci. USA 92: 9737-9741 (1995)). The established function of TRX-G and PC-G genes in chromatin modulation in Drosophila led to the suggestion that the PHD finger is involved in chromatin-mediated transcriptional control. Recent data provide evidence that PHD finger proteins are associated with chromatin remodelling complexes (Bochar D.A. et al. Proc.Natl.Acad.Sci. USA 97: 1038-1043 (2000)) or contribute to histone acetylation (Loewith R. et al. Mol.Cell.Biol. 20: 3807-3816 (2000)). Based on the position of the unique His residue, the cysteine scaffold of the PHD finger (Cys4-His-Cys3) is clearly distinct from RING fingers (Cys3-His-Cys4) and LIM domains (Cys2-His-Cys5) and from DRIL domains, where two RING finger motifs are closely linked. In contrast to the accumulating knowledge about LIM domains, functional data concerning the PHD finger remain rare (see rev. Halbach T. et al. Nucleic Acids Research 28: 3542-3550 (2000)).

GYMNOS, a recently described member of the SWI2/SNF2 protein family in plants (22), also contains a PHD finger and takes part in the control of development. The second PHD finger motif of Drosophila dMI-2 protein (a reference for animal counterparts) shares high sequence conservation to known plant PHD fingers. Due to the similarity to the Drosophila MI-2 gene, GYMNOS has been implicated in chromatin modulation. While the PHD finger is an isolated motif in GYMNOS, the characteristic Cys4-His-Cys3 scaffold in PHDf-HD plant genes is embedded in a large region. This region shares 60% identical residues between seven genes of different plant species and is more highly conserved than the HD (40%). This conservation suggests that the PHD finger is part of a larger functional unit. When combined with a leucine zipper in the surrounding conserved 180 amino acid region in the PHDf-HD proteins, PHD finger activity is masked and silenced. The leucine zipper upstream of the PHD finger mediates interactions with helix 4 of plant 14-3-3 proteins, thus identifying PHDf-HD proteins as potential targets of 14-3-3 signalling pathways. The 14-3-3 family of multifunctional proteins is highly conserved between animals, plants and yeast. Due to the dimeric nature of 14-3-3 proteins and their capacity to form homo- and heterodimers, members of the 14-3-3

protein family function as scaffolds promoting association of protein complexes. 14-3-3 proteins are involved in various signalling pathways that include, for example, Raf, BAD, Bcr/Bcr-Abl, KSR (kinase supressor of Ras), PKC, PI-3 kinase and cdc25C phosphatase. Others enter the nucleus and are associated with DNA-binding complexes. Recent data even indicate contacts to TBP, TFIIB and the human TBP-associated factor hTAF(II)32 (for rev.see Halsbach T., supra).

Recently PHD finger has been shown to activates transcription in yeast, plant and animal cells. Transcriptional activation in animal cells (in the zebrafish embryo as a test system) tested for different PHD fingers seems to be a general feature of the PHD finger motif in eukaryotic cells.

It remains to be elucidated whether the PHD finger directly interacts with a component of the transcription initiation complex or if its positive effect on transcription is mediated via auxiliary protein interactions. Both assumptions, however, involve PHD finger-mediated protein-protein interactions. Surrounding sequences may interfere sterically with accession of the PHD finger and its exposure could eventually depend on binding of a protein partner.

The PHD finger containing proteins appear to be involved in human diseases. Studies on the AIRE gene from humans (Nagamine K. et al. Nat.Genet. 17: 393-398 (1997), Scott H.S. et al. Mol.Endocrinol. 12: 1112-1119 (1998)) have shed more light on the importance of this motif, since all clinically significant mutations in the AIRE gene coincide with alteration in two PHD fingers, resulting in the rare autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED). The presence of PHD fingers in genes up-regulated in leukaemia, associated with the autoimmune disease APECED or participating in euchromatin to heterochromatin modulation, like the TRX-G or PC-G genes, indicates that this motif may be involved in a variety of important cellular events including developmental disorders, tumors and immune diseases. For exemple, the role of a chromatin structure remodelling in cancer metastasis and tissue carcinogenesis is well documented (Zhang Y. et al. Cell 16: 279-289 (1998); Klugbauer S. and Rabes H.M. Oncogene 29: 4388-4393 (1999)).

It is believed that the protein of SEQ ID NO: 388 or part thereof is a zinc binding protein, preferably able to bind nucleic acids, more preferably a transcription factor. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 388 from positions 329 to 339. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 388 having any of the biological activity described herein.

In one embodiment of the invention, the protein of the invention, or part thereof, or derivative thereof, may be used to a subject to diagnose developmental disorders and/or cell proliferative disorders linked to dysregulation of gene expression mediated by the PHD-finger domain of the protein of the invention. Such disorders include but are not limited to, renal tubular acidosis, anemia, Cushing's syndrome, achondroplastic dwarfism, epilepsy, gonadal dysgenesis, hereditary neuropathies such as Charcot-Marie-Tooth disease and neurofibromatosis, hypothyroidism, hydrocephalus, seizure disorders such as Syndenham's chorea and cerebral palsy, spinal bifida, and congenital glaucoma, cataract, sensorineural hearing loss, benign tumors, and cancers such as adenocarcinoma; leukemia; melanoma;

lymphoma; sarcoma; and cancers of the bladder, colon, liver, brain, small intestine, large intestine, breast, ovary, kidney, lung, and prostate. Diagnosis may be performed using nucleic acids or antibodies able to detect the expression of the protein of the invention using any technique known to those skilled in the art including Northern blotting, RT-PCR, immunoblotting methods immunohistochemisty, enzyme-linked immunosorbant assay (ELISA) described herein. Quantities of the protein of the invention expressed in subject samples, control and disease from biopsied tissues or body fluids or cell extracts taken from patients are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease.

In another embodiment, antagonists or inhibitors of the protein of the invention or part thereof may be administered to patients to treat and/or prevent the above referred disorders. Antagonists or inhibitors of transcriptional activators may indeed be used to suppress transcriptional activation in tumor cells. Such antagonists and/or inhibitors may be antibodies specific for the protein of the invention that can be used directly as an antagonist, or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the protein of the invention. Neutralizing antibodies, (i.e., those which inhibit protein-protein interactions) are especially preferred for therapeutic use. Other methods to inhibit the expression of the protein of the invention include antisense and triple helix stategies as described herein. Other antagonists or inhibitors of the protein of the invention may be produced using methods which are generally known in the art, including the screening of libraries of pharmaceutical agents to identify those which specifically bind the protein of the invention. The protein of the invention, or part thereof, preferably its functional or immunogenic fragments, or oligopeptides related thereto, can be used for screening libraries of compounds in any of a variety of drug screening techniques. The fragment employed in such screening may be free in solution, affixed to a solid support, borne on a cell surface, or located intracellularly. The formation of binding complexes, between the protein of the invention, or part thereof, or derivative thereof, and the agent being tested, may be measured. Another technique for drug screening which may be used provides for high throughput screening of compounds having suitable binding affinity to the protein of the invention as described in published PCT application WO84/03564.

Protein of SEQ ID NO: 394 (internal designation 157-17-2-0-C1-CS)

The protein of SEQ ID NO: 394 encoded by the extended cDNA SEQ ID NO: 153 contains a myc-type, helix-loop-helix dimerization domain (Prosite PS00038) from amino acid position 13 to 28 and has no adjacent basic domain. Using the Schiffer-Edmundson helical wheel diagram (Schiffer et al. (1967) Biophys.J. 7:121-135), a hypothetical amphipatic alpha helix is predicted between position 53 and position 68. Three hydrophobic amino acids, Val 55, Phe59 and Ile63, are aligned on the same side of the helix to present a hydrophobic interaction surface and three hydrophilic residues (Tyr53, Gln62 and Ser64) are presented on the other side of helix. There is no Proline residue within the stretch to disrupt the continuity of the alpha helix. Thus, these structural

features in the protein of the invention indicates that this protein could be a novel member of the nonbasic "helix-loop-helix" subfamily (HLH) of transcription regulator.

The helix-loop-helix (HLH) family of transcriptional regulators is involved in the control of different cellular differentiation phenomenon such as neurogenesis, haematopoiesis, myogenesis and angiogenesis. The HLH proteins are found in all eukaryotic organisms ranging from yeast saccharomyces cerevisiae to human (Reviewed by Massari ME and Murre C. (2000) Molecular and Cellular Biology, 20 (2):429-440). The HLH proteins bind DNA as dimers, and different members of HLH family bind either as homodimers or as heterodimers with other members of the family. The presence in a cell of a large repertoire of distinct complexes that can bind to a particular DNA sequence element suggests that competition for DNA binding may play a regulatory role.

Members of the helix-loop-helix (HLH) family of transcriptional regulation proteins share a common structural element, i.e. a stretch of 40-50 amino acids containing two short amphipathic alpha-helices separated by a linker region (the loop) of varying length (Murre C et al. (1989) Cell 56:777-783). This element was initially identified as a region of homology among c-myc, the muscle determination gene MyoD (Davis RL et al. (1987) Cell 51:987-1000) and the Drosophila achaete-scute complex (AS-C) involved in neural determination (Villares R. and Cabrera CV (1987) Cell 50:415-424). The HLH proteins form both homodimers and heterodimers by means of interaction between the hydrophobic residues on the corresponding faces of the two helices to give a parallel four-helix bundle structure (Adrian R et al. (1993) Nature, 363:38-45; Ellenberger T et al. (1994) Genes Dev. 8:970-980). The alpha helical regions are usually 15-16 amino acids long with hydrophobic residues at every third and fourth position, and each helix contains several conserved residues (Murre C et al. (1989) Cell, 56:777-783; Benezra R. et al. (1990) Cell, 61:49-59).

The HLH protein family is subdivided into two major groups: the so-called "bHLH" and "non basic HLH" subfamilies. Proteins of the bHLH family contain a conserved highly basic region immediately N-terminal to the first helix (known as bHLH structure), and mutagenesis experiments on MyoD protein confirm that this region is responsible for sequence-specific binding to the "E-box", a consensus DNA motif for bHLH proteins (Davis RL. et al. (1990) Cell, 60: 733-746). A dimeric bHLH protein (either homodimeric or heterodimeric but in which both subunits contains a basic region) are able to bind to DNA. In general, the bHLH proteins fall into two categories: Class A consists of proteins that are ubiquitously expressed, including mammalian E12/E47 and fly da whereas the class B consists of proteins that are expressed in a more tissue-specific manner, including mammalian MyoD and fly AC-S. In most cases, the tissue-specific bHLH proteins preferentially heterodimerize with ubiquitous partners.

The non basic HLH subfamily contains proteins lacking a basic region unable to bind to DNA but that could form homo- or heterodimers through their HLH motif. Indeed, heterodimeric complexes between non basic HLH and bHLH proteins fail to bind to DNA and negatively modulate the bHLH proteins-mediated transcription activation. This phenomenon was first

demonstrated in a MyoD/Id regulation model (Benezra R. et al. (1990) Cell, 61:49-59). The MyoD gene product is able to activate previously silent muscle-specific genes when introduced into a large variety of differentiated cell types. MyoD proteins form either homodimers or heterodimers with other bHLH proteins such as E12 or E47, and bind to E-box consensus motif to activate myogenesis. The Id gene, conserved from batracians to mammals (Wilson R et al. (1995) Mech.Dev. 49:211-222; Sawai S et al. (1997) Mech.Dev. 65:175-185; Norton JD et al. (1998) trends in Cell Biology 8:58-65), lacks a basic region adjacent to its HLH motif but is able to specifically dimerize with either MyoD,E12 or E14 and has been shown to subsequently attenuate the heterodimer's ability to bind DNA. Additionally, overexpression of Id inhibits MyoD-dependent gene activation in in vivo transfection experiments. Id proteins may function either to repress directly the activity of tissue-restricted bHLH proteins by rendering them non-functional or, more likely, to sequestrate the ubiquitous bHLH proteins and preventing them from forming active heterodimers with the tissue-restricted bHLH (Review by Norton JD et al. (1998) trends in Cell biology 8:58-65).

The possibility that the Id protein behaves as a dominant-negative regulator to repress MyoD protein activity through the formation of nonfunctional heterodimeric complexes is considerably strengthened by the following findings in Drosophila. In Drosophila, the development of peripheral nervous system is positively regulated by the two structurally related bHLH proteins, AS-C and daughterless (da), since loss of either activity results in loss of sensory organ development. The extramacrochaetae Emc product belonging to the non basic HLH subfamily was shown to antagonize the activity of AS-C and da. through the formation of nonfunctional heterodimers with the bHLH proteins (Hillary M et al. (1990) Cell,61:27-38; Garrell J et al. (1990) Cell 61,39-48).

Human Id genes including human Id1, Id2, Id3 and Id4 have been identified and localized (Review by Norton JD et al. (1998) trends in Cell Biology 8:58-65). The bHLH proteins and Id proteins are thought to be involved in the regulation of apoptosis. Differentiation and development of T- and B-lymphocytes in immune system are positively regulated by the combination of ubiquitous E proteins and lymphocyte-restricted bHLH proteins. Disruption in gene expression from either class results in severe perturbation of T- and B-lymphocyte development (Bain G et al. (1997) Mol Cell Biol 17:4782-4791; Zhuang et al. (1996) Mol Cell Biol 16:2898-2905). Cell-arrested T thymocytes undergo a massive apoptosis when Id1 gene is overexpressed (Kim D (1999) Mol Cell Biol 19(12):8240-53). Overexpression of Id1 gene product also results in apoptosis in neonatal and adult cardiac myocytes in culture (Tanaka K et al. (1998)J Biol Chem 273(40) 25922-25928).

Id1 and Id3 proteins are also required to support angiogenesis. Quiescent adult endothelial cells express minimal level of the Id proteins, whereas Id expression is upregulated in angiogenic endothelial cells. Partial loss of these proteins in $Id1^{+/-}Id3^{-/-}$ double knockout mice impairs

angiogenesis, resulting in the resistance to tumour growth (Lyden D et al. (1999) Nature 401:670-677). In addition, a significant overexpression of mRNA and protein levels of Id1, Id2 and Id3 has been found in patients with pancreatic cancer (Maruyama H et al. Am J Pathol (1999) 155(3):815-822) A correlation of Id1 gene upregulation and aggressive phenotype of human breast cancer cells has also been reported (Lin CQ et al. (2000) Cancer Res 60(5):1332-40).

Thus, identification and cloning of members of the HLH family, and especially of the non basic HLH subfamily, is necessary to enrich our knowledge about the biological importance of the HLH transcription factors network and further more to provide insights and tools in disorders linked to dysregulation of the HLH-mediated transcription.

It is believed that the protein of SEQ ID NO: 394 or part thereof plays a role in the regulation of transcription activation, probably as a member of the HLH family, preferably of the non basic HLH subfamily. More particularly, the protein of the invention is thought to be able to antagonize the activity of members of the bHLH family through the formation of heterodimers. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 394 from positions 13 to 28, from positions 53 to 68, and from positions 13 to 68. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 394 having any of the biological activity described herein.

The dimerization ability of the protein of the invention or part thereof which is characteristic of the HLH family may be assayed using any of the assays known to those skilled in the art. For example, interacting protein partners, especially members of the bHLH subfamily, may be identified using screening of cDNA expression libraries as described for the identification of some HLH transcription factors such as E12 and E47 (Murre C et al. (1989) Cell 56:777-783), Max (a Myc binding factor) (Elizabeth M et al. (1991) Science 251:1217) as well as Id (Benezra C et al. (1990) Cell 61:49-59). Alternatively, the helix-loop-helix motif in the protein of the invention could be used by those skilled in art as a "bait protein" in a well established yeast double hybridization system to identify its interacting protein partners in vivo from cDNA library derived from different tissues or cell types of a given organism. Alternatively, the protein of the invention or part thereof could be used by those skilled in art in mammalian cell transfection experiments. When fused to a suitable peptide tag such as $[His]_6$ tag in a protein expression vector and introduced into culture cells, this expressed fusion protein can be immunoprecipitated with its potential interacting proteins by using anti-tag peptide antibody. This method could be chosen either to identify the associated partner or to confirm the results obtained by other methods such as those just mentioned.

An object of the invention relates to compositions and methods using the protein of the invention or part thereof to dysregulate gene transcription, preferably transcription mediated by HLH regulators either in vitro or in vivo, through overexpression of the protein of the invention using any means known to those skilled in the art.

The protein of the invention or part thereof could be used to induce apoptosis of specific cell-type under either physiological or pathological conditions. In a preferred embodiment, the apoptosis active polypeptide is added to an in vitro culture of mammalian cells in an amount effective to induce apoptosis. In another preferred embodiment, the apoptosis active polypeptide is expressed under the control of a promoter which may be activated under precise conditions. In particular, such conditional expression of an apoptosis-active polypeptide upon demand may be very useful to get rid of cells that have become unwanted, for example in applications where such cells have been used in a cellular therapy goal and have become useless. Another example of application is the case of expression under the control of a promoter that becomes active after infection by a given microorganism, thus resulting in the death of the infected cells only. Furthermore, the protein of the invention or part thereof may be useful in the diagnosis, the treatment and/or the prevention of disorders in which apoptosis is beneficial, including but not limited to disorders linked to abnormal cellular proliferation such as those described below.

In another embodiment, the protein of the invention or part thereof can be used to diagnose, treat and/or prevent disorders linked to overexpression of HLH proteins, such as cancer and other disorders relating to abnormal cellular differentiation, proliferation, or degeneration, including hyperaldosteronism, hypocortisolism (Addison's disease), hyperthyroidism (Grave's disease), hypothyroidism, colorectal polyps, gastritis, gastric and duodenal ulcers, ulcerative colitis, and Crohn's disease, neurodegenerative disorders such as Parkinson's or Alzheimer's diseases using any methods and/or techniques described herein. In addition, the protein of the invention or part thereof may be used to evaluate the disease progression and the clinical treatment efficiency. The protein of the invention or part thereof could also be used a molecular target for anti-angiogenesis drug design. Inhibition of protein expression could be achieved by many means known to those skilled in the art including those described in the present application. For example, an antisense nucleotide or triple helix strategy could be developed to block the protein synthesis. Alternatively, the expressed protein of the invention might be neutralized by using specific monoclonal antibody using techniques known to those skilled in the art including those described in Peverali FA et al (1994) EMBO J. 13:4291-4301; Barone MV et al. (1994) Proc.Natl.Acad.Sci.USA 91:4985-4988; and Haza ET et al. (1994) J.Biol.Chem. 269:2139-2145.

Protein of SEQ ID NO: 466 (internal designation 184-4-2-0-D3-CS)

The protein of SEQ ID NO: 466 overexpressed in liver and encoded by the cDNA of SEQ ID NO: 225 displays a Zinc finger motif of RING type (C3HC4) (Pfam signature from positions 41 to 81, Prosite signature from positions 56 to 65) and a B-box zinc finger motif (pfam signature from positions 110 to 153). In addition, the protein of the invention is predicted to have a nuclear localization.

It is believed that the protein of SEQ ID NO: 466 or part thereof is a zinc binding protein, preferably able to bind nucleic acids, more preferably a transcription factor. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 466 from positions 41 to 81 (Ring Zinc finger protein), and from 110 to 153 (B-Box domain). Other preferred polypeptides of the invention are fragments of SEQ ID NO: 466 having any of the biological activity described herein.

Protein of SEQ ID NO: 267 (internal designation 116-111-1-0-H9-CS)

The protein of SEQ ID NO: 267 encoded by the extended cDNA SEQ ID NO: 26 exhibits an Emotif zinc finger domain, C2H2 type, from positions 185 to 202, and is thought to be localized in the nucleus.

It is believed that the protein of SEQ ID NO: 267 or part thereof is a zinc binding protein, preferably able to bind nucleic acids, more preferably a transcription factor. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 267 from positions 185 to 202. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 267 having any of the biological activity described herein.

Protein of SEQ ID NO: 277 (internal designation 160-103-1-0-F11-CS)

The protein of SEQ ID NO: 277 encoded by the extended cDNA SEQ ID NO: 36 exhibits a pfam DHHC zinc finger domain from positions 140 to 204.

It is believed that the protein of SEQ ID NO: 277 or part thereof is a zinc binding protein, preferably able to bind nucleic acids, more preferably a transcription factor. Preferred polypeptides of the invention are polypeptides comprising the residues of SEQ ID NO: 277 from positions 140 to 204. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 277 having any of the biological activity described herein.

Protein of SEQ ID NO: 272 (internal designation 145-25-3-0-B4-CS)

The protein of SEQ ID NO: 272 encoded by the extended cDNA SEQ ID NO: 31 shows homology with numerous zinc binding proteins. In addition, the protein of the invention exhibits the pfam RING zinc finger signature from positions 87 to 129. The protein of SEQ ID NO: 272 has a variant, i.e. the protein of SEQ ID NO: 273 encoded by the extended cDNA SEQ ID NO: 32 and thought to have the same function and utilities.

It is believed that the protein of SEQ ID NO: 272 or part thereof is a zinc binding protein, preferably able to bind nucleic acids or proteins, more preferably a transcription factor. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 272 from positions 87 to 129. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 272 having any of the biological activity described herein.

Hydrolases and inhibitors

The invention relates to compositions and methods using proteins of the invention havinf an hydrolytic activity, herein referred to as HYP, such as the ones described in this section and those containing an hydrolytic domain as shown on Table VI, or parts thereof, preferably fragments comprising an hydrolytic domain, or derivative thereof.

The invention relates to methods and compositions using HYP or a fragment thereof to hydrolyze one or several substrates, alone or in combination with other substances.. For example, the protein of the invention or part thereof is added to a sample containing the substrate(s) in conditions allowing hydrolysis, and allowed to catalyze the hydrolysis of the substrate(s). Hydrolyzed substrates are then detected using standard methods known to those skilled in the arts. The protein of the invention or part thereof can also be added to samples as a "cocktail" with other hydrolytic enzymes, such as other peptidases, for example to decontaminate surgical instruments using methods described in US patent 5,489,531. The advantage of using a cocktail of hydrolytic enzymes is that one is able to hydrolyze a wide range of substrates without necessarily knowing the specificity of each enzyme. Using a cocktail of hydrolytic enzymes also protects a sample from a wide range of future unknown contaminants from a vast number of sources. Alternatively, HYP or part thereof may be bound to a chromatographic support, either alone or in combination with other hydrolytic enzymes, using techniques well known to those skilled in the art, to form an affinity column to remove the substrate. Immobilization facilitates removal of the enzyme from the batch of product and subsequent reuse of the enzyme.

Immobilization of the enzyme or part thereof can accomplished, for example, by adding a cellulose-binding domain to the protein through the modification of the DNA sequence coding for the protein or part thereof. One of skill in the art will understand that other methods of immobilization could also be used and are described in the available literature. Alternatively, the same methods may be used to identify new substrates.

In another embodiment, HYP or part thereof may be used to identify or quantify the amount of a given substrate in a biological sample. In a preferred embodiment, HYP of part thereof is catalytically inactived , i.e. capable of binding but not hydrolyzing a given substrate, using any of the methods known to those skilled in the art including those which produce a mutant enzyme, a recombinant-enzyme, or a chemically inactivated enzyme. The catalytically inactive protein of the invention is then incubated with an aliquot of a biological sample under conditions suitable for binding of the inactive enzyme to the substrate. Then, the bound enzyme is detected to assess the presence or amount of the eubacteria in the biological sample. In another preferred embodiment, HYP or part thereof is used in assays and diagnostic kits for the identification and quantification of substrates in a biological sample. These assays can be based for example, on standard enzyme-linked immunosorbant assays (ELISA) or any other technique known to those skill in the artIn addition, HYP or part thereof may be used to identify, e.g. using screens based on standard assays

such as those described above, inhibitors of the enzyme for mechanistic and clinical applications. Such inhibitors may then be used to identify or quantify HYP in a sample, and to diagnose, treat or prevent any of the disorders where the protein's activity is undesirable and/or deleterious.

Protein of SEQ ID NO:400 (internal designation 160-54-1-0-F7-CS)

The protein of SEQ ID NO:400, encoded by the cDNA of SEQ ID NO:159, exhibits two putative transmembrane domains encompassing amino-acids 50-70 and 127-147 as predicted by the software TopPred II (Claros and von Heijne, *CABIOS applic. Notes*, 10 :685-686 (1994)). It also diplays the Prosite carboxypeptidase zinc-binding region signature PS00133 at positions 117-127. It is predicted by the psort software (see Nakai K and Horton P, Trends Biochem Sci. 1999 Jan;24(1):34-6) to localize to the nucleus with a high probability (73.9%). Finally it is specifically expressed in fetal brain and shows no homology to previously known proteins.

Carboxypeptidase enzymes hydrolyze the terminal amino acid of a protein or peptide. A novel family of carboxypeptidases, localized in the nucleus and with a carboxypeptidase-dependant transcriptional activity, has emerged only recently. Its first member, AEBP1, was previously identified as a 3T3 preadipocyte factor implicated in the repression of the aP2 gene expression. AEBP1 stands for "AE-1 Binding Protein," where AE-1 is a regulatory element of the adipose P2 gene (aP2), a gene involved in triglyceride metabolism and activated in adipocytes. Its own expression is abolished during adipocyte differentiation (He GP et al., Nature 378:92-96(1995)). AEBP1 was subsequently shown to play a similar role in the differentiation of osteoblastic cell lines (Ohno I et al., Biochem Biophys Res Commun. 1996 Nov 12;228(2):411-4) and vascular smooth muscle cells (Layne MD et al., J. Biol. Chem. 273:15654-15660(1998)). It was proposed that AEBP1 acts as a negative transcription factor by cleaving proteins involved in transcription, a new feature in transcription regulation. Recent evidence further suggests that its transcriptional activity is itself attenuated by binding to G-proteins subunits (Park JG et al., EMBO J. 1999 Jul 15;18(14):4004-12) and stimulated by DNA binding (Muise AM and Ro HS, Biochem J. 1999 Oct 15;343 Pt 2:341-5).

It is believed that the protein of SEQ ID NO:400 plays a role in cell signaling, nuclear transcriptional activity and in the differentiation of several cell types, especially those found in the developing brain (including but not limited to neurons). Preferred polypeptides of the invention are polypeptides having any of the biological activities described herein.

One embodiment of the present invention relates to compositions and methods using the protein of the invention or part thereof as a marker for specific cell compartments (especially the nucleus) and/or tissue types (especially fetal brain). For example, the protein of the invention or part thereof may be used to generate specific antibodies which would in turn allow the visualization of nuclear structures by methods well-known to those of skill in the art. In a similar fashion, antibodies raised against the protein of the invention may be used to identify particular

developmental stages (fetal for instance) and/or given tissue types (brain for instance), as the protein of the invention is specifically expressed in brain tissues at a fetal stage. Antibodies and antiserum can also be used to inhibit undesirable carboxypeptidase activities in *in vitro* experiments and cell cultures, as well as in biological samples and in vivo. Alternatively, quantitative analysis or detection of the protein of the invention, or of nucleic acids encoding the protein, can be carried out by any other technique known to those skilled in the art.

In another embodiment, the protein of the invention may be used to target heterologous compounds (polypeptides or polynucleotides) to the developing brain and/or the cell nucleus. For instance, a chimeric protein composed of the protein of the invention recombinantly or chemically fused to a protein or polynucleotide of therapeutic interest would allow the delivery of the therapeutic protein/polynucleotide specifically to the above-mentioned cellular/tissue targets (nucleus, fetal brain).

In another embodiment, the present invention relates to methods and compositions using the protein of the invention or a fragment thereof to hydrolyze one or several substrates, alone or in combination with other substances. The ability of the present protein to hydrolyze any particular substrate can easily be determined by carrying out a hydrolysis reaction using standard assay techniques such as the ones decribed by Slusher et al. (Slusher et al. – Prostate – 2000, 44(1): 55-60) or any other technique well known to those skilled in the art. Potential substrates are any substance containing a peptide bond, more specifically a C-terminal peptide bond. Such substances include, but are not limited to, polypeptides, folic acid and its analogues (e.g. methotrexate). For example, the protein of the invention or part thereof is added to a sample containing the substrate(s) in conditions allowing hydrolysis, and allowed to catalyze the hydrolysis of the substrate(s). Hydrolyzed substrates are then detected using standard methods known to those skilled in the art.

In a preferred embodiment, the protein of the invention or part thereof may be used to modulate cellular transcriptional activity, thereby modulating cellular differentiation. Specifically, as nuclear carboxypeptidases play a role in inhibiting transcription associated with differentiation, then an increase in the activity or expression of the protein can be used to inhibit differentiation. The ability to inhibit differentiation has a number of uses, for example during the cultivation of undifferentiated pluripotent cells to maintain the cultured cells in an undifferentiated state until the need for a given cell type arises (in cases of grafts for instance). The level of the protein activity or expression can be increased in any of a number of ways, including by introducing a polynucleotide encoding the protein into cells, by administering the protein itself to cells, or by administering to cells a compound that increases protein activity or expression. Alternatively, the protein of the invention can be inhibited, thereby enhancing cellular differentiation. The ability to promote differentiation has many uses, including in the treatment or prevention of cancer, as cancer cells are often in a relatively undifferentiated state, and cellular differentiation typically accompanies by growth arrest.

In another embodiment, the protein of the invention or part thereof may be used to diagnose, treat and/or prevent disorders where the presence of substrates, for example excess proteins or peptides, is undesirable or deleterious. Such disorders include but are not limited to, cancer, neurodegenerative disorders such as Parkinson's and Alzheimer's diseases, and diabetes. In another embodiment, the protein of the invention or part thereof may be used to identify or quantify the amount of a given substrate (e.g. a peptide, folic acid, or methotrexate) in a biological sample. In a preferred embodiment, the protein of the invention or part thereof is used in assays and diagnostic kits for the identification and quantification of substrates in a biological sample.

In a most preferred embodiment, the protein of the invention or part thereof can be used in cancer chemotherapies in rescue therapy following toxic high dose methotrexate regimes. Many carboxypeptidases can cleave the C-terminal glutamate moiety from folic acid and its analogues, such as methotrexate. The key role of reduced folates as coenzymes in many biological pathways including those leading to DNA synthesis via the pyrimidines and purines, has made folic acid a target molecule for chemotherapy. Tumor cells grow rapidly and have a high rate of nucleic acid synthesis. Depletion of folic acid has cytotoxic effects, primarily in replicating tissues, and can inhibit growth of tumors with high folic acid requirements. Many carboxypeptidases can directly deplete folate by hydrolytic removal of its glutamate moiety. In cancer chemotherapy, methotrexate (4-amino-N^{10}-methyl-pteroyl-glutamate) is commonly used to deplete the pool of reduced folates by inhibiting dihydrofolate reductase (DHFR), which catalyses the reduction of folates into biologically active tetrahydrofolate form, essential in the biosynthesis of all folate coenzymes. Thus, the protein of the invention or part thereof could be used in rescue therapy following toxic high-dose regimes such as described by Widemann et al. (Widemann B. et al. – Proc. Am. Assoc. Cancer Res. – 1995, 36, p232) and Chabner et al. (Chabner B. et al. – Nature – 1972, 239, p395-397), which disclosures are hereby incorporated by reference in their entity. The basis of this strategy is that hydrolysis of methotrexate produces 4-amino-N^{10}-methyl-pteroate that is about 100 times less active as an inhibitor of DHFR.

In another preferred embodiment, the protein of the invention or part thereof can be used in an enzyme/prodrug strategy to treat a number of pathologies, especially those treated with drugs associated with severe side effects, including, but not limited to, autoimmune diseases and chronic inflammatory diseases such as rheumatoid arthritis, and cancer chemotherapy. These side effects can be mainly explained by the fact that the in vivo selectivity of the drugs used is too low (for example, the inadequate selectivity between tumor and normal cells of most anticancer drugs is well known and their toxicity to normal tissues is dose limiting). In the first phase of one example of such a protocol, a conjugate of the protein of the invention or part thereof and an antibody to a tissue specific antigen (for example, tumor specific antigens in the case of cancer chemotherapy) is administered. After a delay to allow residual enzyme conjugate to be cleared from the blood, a relatively non-toxic compound is administered to the patient. This non-toxic compound is a

substrate of the protein of the invention, and is converted by the protein into a substantially more toxic compound. Thus, because of the previous, targeted administration of the protein of the invention, when the non-toxic compound is administered, the toxic compound is only produced in the vicinity of the cells targeted by the fusion protein. This two-phase approach has been termed antibody-directed enzyme-prodrug therapy (ADEPT), this approach is reviewed by Melton et al. (Melton R. et al. – J. Natl. Cancer Inst. – 1996, 88, p153-165). Alternatively the first phase can be replaced by a gene therapy approach resulting in the de novo synthesis of the protein of the invention or part thereof by cells from the targeted tissue, this has been termed gene-dependent enzyme/prodrug therapy (GDEPT). Another advantage of these 2 approaches (ADEPT and GDEPT) is that a single enzyme molecule is capable of activating many prodrug molecules.

Protein of Seq Id No: 242 (internal designation 119-003-4-0-C2-CS)

The protein of SEQ ID No: 242, encoded by the cDNA of SEQ ID No: 1, is homologous to proteins of the M20 metallopeptidases family (EC 3.4.17.X). The protein of the invention is over-expressed in the spinal cord and the brain.

The M20 metallopeptidase family of proteins are all peptidases (i.e. enzymes able to hydrolyze peptide bonds) furthermore they are all exopeptidases, which means that they can hydrolyze the terminal amino acid of a protein or peptide. Members of the M20 peptidase family are glutamate carboxypeptidases, which are capable of releasing the C-terminal glutamate residue, by hydrolysis, from a wide range of N-acyl groups, including peptidyl, aminoacyl, benzoyl, benzyloxycarbonyl, folyl, and pteroyl groups, and physiologically are involved in the catabolism of proteins. M20 carboxypeptidases are either monomeric or homodimeric (i.e. 2 identical proteins assembled to from the enzyme). In order to be active, metallopeptidases must be associated with a metallic cofactor (either Zinc or Cobalt depending on the enzyme). The most studied carboxypeptidase of the M20 family is carboxypeptidase G2 (CPG2) (EC 3.4.17.11), a bacterial enzyme from Pseudomonas sp. (strain RS-16). CPG2 is a dimeric Zinc carboxypeptidase that cleaves the C-terminal glutamate moiety from a number of molecules.

The protein of SEQ ID No: 242 includes the pfam signature for M20 peptidase (position 107 to 451). The protein of SEQ ID No: 242 also includes a number of amino acids that are conserved throughout the M20 protease family especially those that interact with the metal cofactor. Preferred polypeptides of the invention are polypeptides of SEQ ID No: 242 that include the highly conserved amino acids: 133, 135, 149, 163, 200, 201 and/or 262, which are present in over 80% of the members of the M20 peptidase family, and/or amino acids 139, 157, 162, 16, 367 and/or 377, which are present in over 60% of the members of the M20 peptidase family. Of particular interest are amino acids 133, 166, 201 and 262, which by homology are probably involved in the interaction with the metal cofactors. Thus it is believed that the protein of SEQ ID No: 242 or part thereof is a peptidase, preferably a carboxypeptidase, more preferably a metallocarboxypeptidase of the M20

family. Other preferred polypeptides of the invention are any fragments of SEQ ID No: 242 having any of the biological activities described herein.

Determination of carboxypeptidase activity on specific substrates can easily be obtained by carrying out the hydrolysis using standard assay techniques such as the ones decribed by Slusher et al. (Slusher et al. – Prostate – 2000, 44(1): 55-60) or any other technique well known to those skilled in the art. Potential substrates are any substance containing a peptide bond, more especially C-terminal peptide bonds, and even more specifically, C-terminal glutamate. Such substances include but are not limited to peptides, folic acid and its analogues (e.g. methotrexate).

In an embodiment the protein of the invention or part thereof could be used to develop assay tools to identify brain and spinal cord tissue since the protein of the invention is overexpressed in these tissues.

In still another embodiment, the protein of the invention or part thereof may be used to diagnose, treat and/or prevent disorders where the presence of substrates, for example excess proteins, is undesirable or deleterious. Such disorders include but are not limited to, cancer, neurodegenerative disorders such as Parkinson's and Alzheimer's diseases, and diabetes. In a most preferred embodiment, the protein of the invention or part thereof can be used in cancer chemotherapies in rescue therapy following toxic high dose methotrexate regimes. Enzymes of the M20 peptidase family can cleave the C-terminal glutamate moiety from folic acid and its analogues, such as methotrexate. The key role of reduced folates as coenzymes in many biological pathways including those leading to DNA synthesis via the pyrimidines and purines, has made folic acid a target molecule for chemotherapy. Tumor cells grow rapidly and have a high rate of nucleic acid synthesis. Depletion of folic acid has cytotoxic effects, primarily in replicating tissues, and can inhibit growth of tumors with high folic acid requirements. Enzymes of the M20 peptidase family can directly deplete folate by hydrolytic removal of its glutamate moiety. In cancer chemotherapy, methotrexate (4-amino-N^{10}-methyl-pteroyl-glutamate) is commonly used to deplete the pool of reduced folates by inhibiting dihydrofolate reductase (DHFR), which catalyses the reduction of folates into biologically active tetrahydrofolate form, essential in the biosynthesis of all folate coenzymes. Thus the protein of the invention or part thereof could be used in rescue therapy following toxic high-dose regimes such as described by Widemann et al. (Widemann B. et al. – Proc. Am. Assoc. Cancer Res. – 1995, 36, p232) and Chabner et al. (Chabner B. et al. – Nature – 1972, 239, p395-397), which disclosures are hereby incorporated by reference in their entity. The basis of this strategy is that hydrolysis of methotrexate produces 4-amino-N^{10}-methyl-pteroate that is about 100 times less active as an inhibitor of DHFR.

In another preferred embodiment, the protein of the invention or part thereof can be used in an enzyme/prodrug strategy to treat a number of pathologies, especially those treated with drugs associated with severe side effects, including, but not limited to, autoimmune diseases and chronic inflammatory diseases such as rheumatoid arthritis, and cancer chemotherapy. These side effects

can be mainly explained by the fact that the in vivo selectivity of the drugs used is too low (for example, the inadequate selectivity between tumor and normal cells of most anticancer drugs is well known and their toxicity to normal tissues is dose limiting). In the first phase of one example of such a protocol, a conjugate of the protein of the invention or part thereof and an antibody to a tissue specific antigen (for example, tumor specific antigens in the case of cancer chemotherapy) is administered. After a delay to allow residual enzyme conjugate to be cleared from the blood, a relatively non-toxic compound is administered to the patient. This non-toxic compound is a substrate of the protein of the invention, and is converted by the protein into a substantially more toxic compound. Thus, because of the previous, targeted administration of the protein of the invention, when the non-toxic compound is administered, the toxic compound is only produced in the vicinity of the cells targeted by the fusion protein. This two-phase approach has been termed antibody-directed enzyme-prodrug therapy (ADEPT), this approach is reviewed by Melton et al. (Melton R. et al. – J. Natl. Cancer Inst. – 1996, 88, p153-165). Alternatively the first phase can be replaced by a gene therapy approach resulting in the de novo synthesis of the protein of the invention or part thereof by cells from the targeted tissue, this has been termed gene-dependent enzyme/prodrug therapy (GDEPT). Another advantage of these 2 approaches (ADEPT and GDEPT) is that a single enzyme molecule is capable of activating many prodrug molecules.

Protein of SEQ ID NO: 401 (internal designation 160-88-3-0-A8-CS.corr)

The protein of SEQ ID NO : 401 encoded by the cDNA SEQ ID NO: 160 is a splicing variant of the hypothetical human palmitoyl-protein thioesterase-2 (PPT2) (E.C. 3.1.2.22) (Genbank accession number AF020543), which is well conserved among eukaryotes (C. elegans and rodents) and exhibits homology with the palmitoyl protein thioesterase-1 (PPT1) (Genbank accession number L42809). The product of the cDNA SEQ ID NO: 160 is shorter than the human PPT2 (280 versus 308 amino acids respectively) with a gap located between the positions 174 and 203 of the protein PPT2. The protein of SEQ ID NO : 401 has a variant, the protein of SEQ ID NO: 402 encoded by the cDNA of SEQ ID NO: 161, thought to have the same functions and utilities.

PPT1 (E.C. 3.1.2.22) is a well-described protein, widely conserved among the murine, rat, bovine and human species (Swissprot accession number P50897). It is a lysosomal enzyme that functions in the removal of fatty acids from modified cysteine residues in proteins undergoing degradation (Hofmann S.L. *et al, Neuropediatrics*, **28**: 27-30 (1997)). For example, PPT1 catalyses the deacylation H-ras and the alpha subunits of heterodimeric G proteins *in vitro* (Camp L.A., *J. Biol. Chem.*, **268**: 22566-22574 (1993) and **269:** 23212-23219 (1994)). Deacylation by PPT1 may be a prerequisite for complete digestion of the modified polypeptides. In fact there is evidence that palmitoylation leads to increased protection against proteolytic digestion. Both the salivary mucus glycoprotein (Slomiany B. L., *Biochem. Biophys. Res. Commun.*, **151**: 1046-1053 (1988),) and chemically acylated bee venom phospholipase A2 (Diaz, R.E., *Biochem. Biophys. Acta*, **830**: 52-58

(1985)) are more resistant to treatment with proteinases than their deacylated forms. Mutations in PPT1 enzyme were shown to underlie the hereditary neurodegenerative disorder, infantile neuronal ceroid lipofuscinosis (Vesa *et al.*, *Nature*, **376**: 584-587 (1995)).

Recently, Soyombo and Hofmann (*J.Biol.Chem*, **272**: 27456-27463, (1997)) described a second lysosomal thioesterase, PPT2, that shares 20% identity with PPT1. The PPT2 enzyme presumably also plays a role in lysosomal thioester catabolism but has a substrate specificity distinct from that of PPT1. While little is known about the substrate specificity of PPT2, the enzyme is highly active against palmitoylated model substrates such as palmitoyl CoA. PPT2 did not hydrolyse the acyl-cysteine bond of the protein substrates routinely used to assay PPT1 such as H-Ras and albumin. This finding suggest that although both enzymes possess intrinsic palmitoyl thioesterase activity, the "leaving group" recognized by the enzymes may differ. One possibility is that PPT2 recognizes palmitoylated protein substrates but that these substrates differ from those recognized by PPT1. A second possibility is that PPT2 recognizes a novel lipid thioester substrate that is not derived from acylated proteins. Aguado *et al.* (*Biochem J.*, **341**:679-689, (1999)) demonstrated that PPT2 is an acyl thioesterase . However they cannot distinguish between esterase (thioesterase) and lipase activity. PPT2 shows very high S-thioesterase activity towards the acyl chains $C_{14:0}$>$C_{16:0}$, moderate activity towards the acyl chains $C_{14:1}$>$C_{20:4}$ ≈ $C_{16:1}$ ≈ $C_{18:0}$ ≈ $C_{12:0}$>$C_{18:2}$ ≈ $C_{18:3}$>$C_{22:1}$ ≈ $C_{18:1}$≈ $C_{20:0}$, low activity towards the acyl chains $C_{10:0}$ and $C_{22:0}$, and no activity towards the acyl chain $C_{24:0}$, $C_{8:0}$, $C_{6:0}$, $C_{4:0}$ and $C_{2:0}$. PPT2 has a broader range of action than PPT1, although both have a preference for long acyl chains (more than 12 or 14 carbons) over shorter acyl chains (less than 12 carbons). Aguado *et al.* (supra) also presented a detailed characterization of PPT2 gene product. The putative 302-residue PPT2 and the protein of the invention contains a hydrophobic leader peptide at the N-terminus (signal peptide with a cleavage site predicted at position 34 of the protein of the invention) suggesting that they are secretory glycoproteins. Both proteins exhibit two motifs located at the N-terminus from positions 108 to 121. One motif is common to triglycerides lipases (from position 110 to 121) and the other one to eukaryotic thiol (Cys) proteases (from positions 108 to 121). Triglyceride lipases are lipolytic enzymes that hydrolyse the ester bond of triglycerides. The most conserved region in all these proteins is centered on a serine residue located in a conserved Gly-Xaa-Ser-Xaa-Gly motif. The PPT2 protein and the protein of the invention contain a cysteine residue (position 115) instead of the first glycine residue in the motif but other lipases with one mismatch in either of the consensus have been described (Blow D., *Nature*, **343**: 694-695 (1990)). In the same region as the lipase motif, PPT2 and the protein of the invention contains a motif common to the active site of eukaryotic thiol (Cys) protease but with a leucine residue (position 113) instead of the glycine at the position 5 of the pattern. In addition, the amino acid sequence of the putative PPT2 shows, at the C-terminus, from positions 171 to 186, a motif common to growth factor and cytokine receptors family, which is not present in the protein of the invention.

Aguado *et al.* (supra) have found that PPT2 is expressed in cells of the immune system as an approximatively 42 kDa protein in cells extracts and supernatants and is transcribed as at least five different transcripts. The PPT2 gene is located in the class III region of the human MHC which contains several genes encoding proteins with potential roles in the immune system and in inflammation. In addition, Aguado *et al.* (supra) showed that very large amounts of PPT2 are secreted. However this is not in disagreement with an intracellular activity because the secreted protein could be internalized into the cell through a receptor and act on target located in an intracellular organelle. This mechanism has been described for the secreted PPT1, which can be internalized into the cell by mannose-6-phosphate receptor to act in the lysosome (Verkruyse and Hofmann, *J.Biol.Chem.*, **271**: 15831-15836, (1996)), and Soyombo and Hofmann (*J.Biol.Chem*, **272**: 27456-27463, (1997)) reported that PPT2 binds to mannose-6 phosphate receptor.

Palmitoylation refers to posttranslational modification of proteins in which the most common fatty acids of the cell (i.e. palmitic, stearic and oleic acids) are attached to the side chain of cysteine residues via high-energy thioester linkages (Bizzozero, O.A. *et al*, *Neurochem.Res.*, **19**: 923-933 (1994); Casey P.J., *Science*, **268**: 221-225 (1995)). At present a large number of proteins of diverse origin, structure and function are known to be modified with these fatty acids that attach them to inner surface of the plasma membrane, where the can function optimally (Casey P.J., *Science*, **268**: 221-225 (1995)). Being anchored to membranes is a process necessary for the diverse cellular functions of these modified proteins, including signal transduction, vesicle transport and maintenance of the cytoarchitecture. Almost every tissue and subcellular organelle contains characteristic set of palmitoylated proteins.

The protein of the invention is overexpressed in brain. In recent years a considerable number of functionally relevant nervous system proteins including ions channels, neurotransmitter receptors, signal transduction components and cell-adhesion molecules have been found to be palmitoylated. Although the nervous system is not an exception to this rule, both the number of modified protein in this tissue and the dynamic nature of protein palmitoylation suggest that this modification is critical for regulating important biological processes and that the addition or removal of the fatty acid serves to regulate the activity of these proteins rather that to define their function.

It is believed that the protein of SEQ ID NO: 401 or part thereof is an hydrolase, preferably acting on ester bonds, more preferably a thiolester hydrolase, even more preferably an acyl-thioesterase which, as such, plays a role in fatty acid metabolism, in cellular vesicle transport and maintenance of the cytoarchitecture, in cellular proteolysis, endocytosis, signal transduction, lysosomal storage, cell proliferation and differentiation, immune and inflammatory response. The enzyme's substrates are compounds preferably containing an ester bond, preferably a thiol ester bond, more preferably an acyl thioester bond. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 401 from positions 108 to 121, and 110 to

121. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 401 having any of the biological activities described herein. The hydrolytic activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in Smith *et al., Biochem.J.*, **212**: 155 (1983), Spencer *et al., J.Biol.Chem.*, **253**: 5922 (1978) and Aguado *et al.* (supra) or in US patents 5,445,942.

In another preferred embodiment, the protein of the invention or part thereof may be used to diagnose, treat and/or prevent disorders where the presence of substrates is undesirable or deleterious. Such disorders include but are not limited to infantile neuronal ceroid lipofuscinosis and lysosomal diseases. For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. For prevention and/or treatment purposes, the expression of protein of the invention may be enhanced using any of the gene therapy methods described herein or known to those skilled in the art.

In addition, the protein of the invention or part thereof may be used to identify inhibitors for mechanistic and clinical applications. Such inhibitors may then be used to identify or quantify the protein of the invention in a sample, and to diagnose, treat or prevent any of the disorders where the protein's hydrolytic activity is undesirable and/or deleterious including but not limited to lysosomal diseases, neurodegenerative disorder such as infantile neuronal ceroid lipofuscinosis, Parkinson's and Alzheimer's diseases, inflammatory and immune disorders including allergies and leukemia.

Another object of the present invention are compositions and methods of targeting heterologous compounds, either polypeptides or polynucleotides to lysosomes by recombinantly or chemically fusing a fragment of the protein of the invention to an heterologous polypeptide or polynucleotide. Preferred fragments are any fragments of the protein of the invention, or part thereof, that may contain targeting signals for lysosomes such as those described in Vitale et al, Mol.Cell.Biol, **20**: 7342-52 (2000), Blagoveshchenskaya et al., J.Biol.Chem., **273**: 2729-37 (1998) and Kornfeld, FASEB J., **1**: 462-8 (1987)). Such heterologous compounds may be used to modulate lysosomal activity. For example, they may be used to induce and/or prevent a lysosomal protein degradation. Moreover, antibodies binding to the protein of the invention or part thereof may be used for detection of the lysosomes using any techniques known to those skilled in the art.

In still another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably brain tissues. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Another embodiment of the present invention relates to methods and compositions using the protein of the invention or part thereof to modify plant lipid composition using any assay known to those skilled in the art including those described by the US patents 5,955,650, 5,945,585 and 5,807,893. Indeed, plant lipids have a variety of nutritional uses and many recent research efforts have examined the role that saturated and unsaturated fatty acids play in reducing the risk of coronary heart disease. In the past, it was believed that mono-unsaturates, in contrast to saturates and poly-unsaturates, had no effect on serum cholesterol and coronary heart disease risk. Several recent human clinical studies suggest that diets high in mono-unsaturated fat and low in saturated fat may reduce the "bad" (low-density lipoprotein) cholesterol while maintaining the "good" (high-density lipoprotein) cholesterol (Mattson *et al., Journal of Lipid Research,* **26**: 194-202 (1985)).

In still another embodiment , the protein of the invention or part thereof may be used in enzyme replacement therapy, due to the ability of cells to take up exogeneously supplied protein and target it to lysosomes (Neufeld E.F., *Annu.Rev.Biochem.* **60**: 257-280(1991), Brady R.O. *et al, J.Inher.Metab.Dis.* **17**: 510-519 (1994)), or in bone-marrow transplantation (Hoogerbrugge P.M. *et al, Lancet,* **345**: 1398-1402 (1995)), as bone-marrow-derived microglial cells are believed to penetrate the blood-brain barrier and may theoretically be able to provide sufficient enzyme to correct the metabolic defect in neurons (Krivit W., *Cell transplant.,* **4**: 385-392 (1995)). The protein of the invention or part thereof may be also used in genetic engineering of transplanted cells (Salvetti A. *et al, Br.Med.J.* 51: 106-122 (1995)) or neural progenitor cell engraftment (Snyder E.Y., *Nature*, 374: 367-370 (1995)) using any technique known to those skilled in the art.

Protein of SEQ ID NO: 254 (internal designation 106-006-1-0-E3-CS)

Angiogenin is a member of the pancreatic Rnase superfamily of proteins. Its mechanism of action is postulated to involve multiple interactions with other proteins through specific regions on the molecular surface of angiogenin. Potential partners of angiogenin include heparin, plasminogen, elastase, angiostatin, actin, and a 170 kDa receptor on the surface of endothelial cells [Strydom, D. J. (1998) Cell. Mol. Life Sci. 54, 811-824].

Angiogenin is required for the process of angiogenesis. Tumor growth requires angiogenesis, and several anti-angiogenic agents have been produced and are currently in the clinical trial stage. It has also been shown that recurrent gastric cancer patients had a much higher serum concentration of angiogenin than primary gastric cancer patients [Shimoyama, S. and Kaminishi, M. (2000) J. Cancer Res. Clin. Oncol. 126, 468-474]. Therefore, angiogenin can be used as a diagnostic marker for the evaluation of cancer aggressiveness or as an early marker for recurrence over a follow-up period.

Angiogenin is a potent inducer of angiogenesis [Fett, J. W.; Strydom, D. J.; Lobb, R. R.; Alderman, E. M.; Bethune, J. L.; Riordan, J. F.; and Vallee, B. L. (1985) Biochemistry 24, 5480-5486]. Angiogenesis is a complex process of blood vessel formation comprising of several separate

but interconnected steps at the cellular and biochemical level including: (i) activation of endothelial cells by the action of an angiogenic stimulus, (ii) adhesion and invasion of activated endothelial cells into the surrounding tissues and migration toward the source of the angiogenic stimulus, and (iii) proliferation and differentiation of endothelial cells to form a new microvasculature [Folkman, J. and Shing, Y. (1992) J. Biol. Chem. 267, 10931-10934; Moscatelli, D. and Rifkin, D. B. (1988) Biochim. Biophys. Acta 948, 67-85].

Angiogenin has been demonstrated to induce most of the individual events in the process of angiogenesis including binding to endothelial cells [Badet, J.; Soncin, F.; Guitton, J.D.; Lamare, O.; Cartwright, T.; and Barritault, D. (1989) Proc. Natl. Acad. Sci U.S.A. 86, 8427-8431], stimulating second messengers [Bicknell, R. and Vallee, B. L. (1988) Proc. Natl. Acad. Sci. U.S.A. 85, 5961-5965], mediating cell adhesion [Soncin, F. (1992) Proc. Natl. Acad. Sci. U.S.A. 89, 2232-2236], activating cell-associated proteases [Hu, G. F. and Riordan, J. F. (1993) Biochem. Biophys. Res. Commun. 197, 682-687], inducing cell invasion [Hu, G-F.; Riordan, J. F.; and Vallee, B. L. (1994) Proc. Natl. Acad. Sci. U.S.A. 91, 12096-12100], inducing proliferation of endothelial cells [Hu, G-F.; Riordan, J. F.; and Vallee, B. L. (1997) Proc. Natl. Acad. Sci. U.S.A. 94, 2204-2209] and organizing the formation of tubular structures from the cultured endothelial cells [Jimi, S-I.; Ito, K-I.; Kohno, K.; Ono, M.; Kuwano, M.; Itagaki, Y.; and Isikawa, H. (1985) Biochem. Biophys. Res. Commun. 211, 476-483]. Angiogenin has also been shown to undergo nuclear translocation in endothelial cells via receptor-mediated endocytosis [Moroianu, J. and Riordan, J. F. (1994) Proc. Natl. Acad. Sci. U.S.A. 91, 1677-1681] and nuclear localization sequence-assisted nuclear import [Moroianu, J. and Riordan, J. F. (1994) Biochem. Biophys. Res. Commun. 203, 1765-1772].

While angiogenesis is a tightly-controlled process under usual physiological conditions, abnormal angiogenesis can have devastating consequences in pathological conditions such as arthritis, diabetic retinopathy and tumor growth. It is now well-established that the growth of virtually all solid tumors is angiogenesis dependent [Folkman, J. (1989) J. Natl. Cancer Inst. 82, 4-6]. Angiogenesis is also a prerequisite for the development of metastasis, since it provides the means whereby tumor cells disseminate from the original primary tumor and establish at distant sites [Mahadevan, V. and Hart, I. R. (1990) Rev. Oncol. 3, 97-103; Blood, C. H. and Zetter B. R. (1990) Biochim. Biophys. Acta 1032, 89-118]. Therefore, interference with the process of tumor-induced angiogenesis can be an effective therapy for both primary and metastatic cancers.

Although originally isolated from medium conditioned by human colon cancer cells (Fett et al. (1985), supra), and subsequently shown to be produced by several other histological types of human tumors [Rybak, S. M.; Fett, J. W.; Yao, Q-Z.; and Vallee, B. L. (1987) Biochem. Biophys. Res, Commun. 146, 1240-1248; Olson, K. A.; Fett, J. W.; French, T. C.; Key, M. E.; and Vallee, B. L. (1995) Proc. Natl. Acad. Sci. U.S.A. 92, 442-446], angiogenin also is a constituent of human plasma and normally circulates at a concentration of 250-360 ng/ml [Shimoyama, S.; Gansauge, F.;

Gansauge, S.; Negri, G.; Oohara, T.; and Beger, H. G. (1996) Cancer Res. 56, 2703-2706; Blaser, J.; Triebl, S.; Kopp, C.; and Tschesche, H. (1993) Eur. J. Clin. Chem. Clin. Biochem. 31, 513-516].

Several inhibitors of the functions of angiogenin have been developed. These include: (i) monoclonal antibodies (mAbs) [Fett, J. W.; Olson, K. A.; and Rybak, S. M. (1994) Biochemistry 33, 5421-5427], (ii) an angiogenin-binding protein [Hu, G-F.; Chang, S-I.; Riordan, J. F.; and Vallee, B. L. (1991) Proc. Natl. Acad. Sci. U.S.A. 88, 2227-2231; Hu, G-F.; Strydom, D. J.; Fett, J. W.; Riordan, J. F.; and Vallee, B. L. (1993) Proc. Natl. Acad. Sci. U.S.A. 90, 1217-1221; Moroianu, J.; Fett, J. W.; Riordan, J. F.; and Vallee, B. L. (1993) Proc. Natl. Acad. Sci. U.S.A. 90, 3815-3819], (iii) the placental ribonuclease inhibitor (PRI) [Shapiro, R. and Vallee, B. L. (1987) Proc. Natl. Acad. Sci. U.S.A. 84, 2238-2241], (iv) peptides synthesized based on the C-terminal sequence of angiogenin [Rybak, S. M.; Auld, D. S.; St. Clair, D. K.; Yao, Q-Z.; and Fett, J. W. (1989) Biochem. Biophys. Res. Commun. 162, 535-543], and (v) inhibitory site-directed mutagenesis of angiogenin [Shapiro, R. and Vallee, B. L. (1989) Biochemistry 28, 7401-7408].

The subject invention provides the protein/polypeptide of SEQ ID NO: 254. The invention also provides biologically active fragments of SEQ ID NO: 254. In one embodiment, the polypeptides of SEQ ID NO: 254 are interchanged with the corresponding polypeptides encoded by the human cDNA of clone 106-006-1-0-E3-CS. "Biologically active fragments" are defined as those peptide or polypeptide fragments having at least one of the biological functions of the full length protein (e.g., stimulation of angiogenesis). Compositions of the protein/polypeptide of SEQ ID NO: 254, or biologically active fragments thereof, are also provided by the subject invention. These compositions may be made according to methods well known in the art.

The invention also provides variants of the protein of SEQ ID NO: 254. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence encoded by SEQ ID NO: 254. Variants according to the subject invention also have at least one functional or structural characteristic of the protein of SEQ ID NO: 254. The invention also provides biologically active fragments of the variant proteins. Compositions of variants, or biologically active fragments thereof, are also provided by the subject invention. These compositions may be made according to methods well known in the art. Unless otherwise indicated, the methods disclosed herein can be practiced utilizing the protein encoded by SEQ ID NO: 254, biologically active fragments of SEQ ID NO: 254, variants of SEQ ID NO: 254, and biologically active fragments of the variants.

Because of the redundancy of the genetic code, a variety of different DNA sequences can encode the amino acid sequence of SEQ ID NO: 254. In a preferred embodiment, SEQ ID NO: 254 is encoded by clone 106-006-1-0-E3-CS. It is well within the skill of a person trained in the art to create these alternative DNA sequences which encode proteins having the same, or essentially the same, amino acid sequence. These variant DNA sequences are, thus, within the scope of the subject invention. As used herein, reference to "essentially the same" sequence refers to sequences that

have amino acid substitutions, deletions, additions, or insertions that do not materially affect biological activity. Fragments retaining one or more characteristic biological activity of the protein encoded by clone 106-006-1-0-E3-CS are also included in this definition.

"Recombinant nucleotide variants" are alternate polynucleotides which encode a particular protein. They can be synthesized, for example, by making use of the "redundancy" in the genetic code. Various codon substitutions, such as the silent changes which produce specific restriction sites or codon usage-specific mutations, can be introduced to optimize cloning into a plasmid or viral vector or expression in a particular prokaryotic or eukaryotic host system, respectively.

In one aspect of the subject invention, SEQ ID NO: 254, and variants thereof, can be used to generate polyclonal or monoclonal antibodies. Both biologically active and immunogenic fragments of SEQ ID NO: 254, or variant proteins, can be used to produce antibodies. Polyclonal and/or monoclonal antibodies can be made according to methods well known to the skilled artisan. Antibodies produced in accordance with the subject invention can be used in a variety of detection assays known to those skilled in the art. The antibodies may be used to agonize or antagonize the biological activity of the protein of SEQ ID NO: 254.

SEQ ID NO: 254 can be used as a marker for individuals at risk for the development or recurrence of tumors. As indicated supra, angiogenin is found at certain levels in normal individuals, normally at concentrations of 250-360 ng/ml. Thus, quantitative immunoassays can be used for the detection of abnormal levels of SEQ ID NO: 254, thereby identifying those individuals at risk for the development of tumors. Alternatively, the subject invention provides antibodies specific for SEQ ID NO: 254, or fragments thereof, which are used in routine immunoassays to screen for the presence or absence of SEQ ID NO: 254, or fragments thereof.

Alternatively, the nucleic acids which encode SEQ ID NO: 254, or fragments thereof, may be used in hybridization assays to detect and/or quantitate the expression of SEQ ID NO: 254. Such hybridization assays are well known to the skilled artisan and can be practiced on a variety of samples, including, but not limited to, tumor cells, biopsied tissues, or normal tissue.

Molecules (see Strydom, D. J., (1998) Cell. Mol. Life Sci. 54, 811-824) that functionally inhibit the action of angiogenin can be used to treat patients with tumors. Because angiogenin is required for the vascularization of tumors, molecules which inhibit the biological activity of angiogenin can be used to reduce tumor vascularization and control tumor growth. Thus, another aspect of the invention provides molecules which inhibit, or reduce, the biological activity of SEQ ID NO: 254. One embodiment provides neutralizing antibodies to inhibit the biological activity of SEQ ID NO: 254. These neutralizing antibodies may be chimeric or humanized, according to methods well known in the art, to minimize the immunogenicity of the molecules when used in patients. Neutralizing antibodies may be used in conjunction with other known therapeutic modalities for the treatment of tumors.

Another embodiment of the invention utilizes the concept that expression of specific genes can be suppressed by oligonucleotides having a nucleotide sequence complementary to the mRNA transcript of the target gene. This suppression occurs by selectively impeding translation and has been termed an "antisense" methodology. In addition, "antigene" or "triplex" methodologies may also suppress expression of genes by using an oligonucleotide which is complementary to a selected site of double stranded DNA, thereby forming a triple-stranded complex to selectively inhibit transcription of the gene. Both "antisense" and "antigene" methodologies can be used to inhibit or reduce the expression of the gene of SEQ ID NO: 254, and thereby provide therapeutic benefit to the patient being treated. Methods of treating individuals using antigene and antisense methodologies are well known to those skilled in the art (see, for example, "Antisense Therapeutics" Agrawal, S. (ed), Humana Press, 1996; Crooke, S. T., and Bennett, C. F. (1996) Annu. Rev. Pharmacol. Toxicol. 36, 107-129; "Prospects for the Therapeutic Use of Antigene Oligonucleotides", Maher, L. J. (1996) Cancer Investigation 14(1), 66-82 each hereby incorporated by reference in its entirety).

As additional examples, U.S. Pat. No. 5,098,890 is directed to antisense oligonucleotides complementary to the c-myb oncogene and antisense oligonucleotide therapies for certain cancerous conditions. U.S. Pat. No. 5,135,917 provides antisense oligonucleotides that inhibit human interleukin-1 receptor expression. U.S. Pat. No. 5,087,617 provides methods for treating cancer patients with antisense oligonucleotides. U.S. Pat. No. 5,166,195 provides oligonucleotide inhibitors of HIV. U.S. Pat. No. 5,004,810 provides oligomers capable of hybridizing to herpes simplex virus Vmw65 mRNA and inhibiting replication. U.S. Pat. No. 5,194,428 provides antisense oligonucleotides having antiviral activity against influenza virus. U.S. Pat. No. 4,806,463 provides antisense oligonucleotides and methods using them to inhibit HTLV-III replication. U.S. Pat. No. 5,286,717 is directed to a mixed linkage oligonucleotide phosphorothioates complementary to an oncogene. U.S. Pat. No. 5,276,019 and U.S. Pat. No. 5,264,423 are directed to phosphorothioate oligonucleotide analogs used to prevent replication of foreign nucleic acids in cells. Each of these patents is hereby incorporated by reference in its entirety.

The subject invention also provides modified/derivatized nucleic acids encoding SEQ ID NO: 254. These include those modifications which increase the stability and/or affinity of these compounds for targets. Phosphorothioate analogs of oligodeoxynucleotides (ODNs), in which nonbridging phosphoryl oxygens in the backbone of DNA are substituted with sulfur ([S]ODNs) are substantially more stable than their native phosphodiester counterparts. Other derivatives, such as those alkylated on sugar oxygen groups, show enhanced target affinity. [S]ODNs possess good biological activity, pharmacology, pharmacokinetics and safety *in vivo* (Agrawal (1996), supra). Successful inhibition of specific gene function has been achieved by targeting various sites on specific mRNA sequences that include the AUG translational initiation codon, 5'-transcriptional

start site, 3'-termination codon and sequences in both the 5' and 3'-untranslated regions. These derivatized nucleic acids can be used in any of the aforementioned methodologies.

Protein of SEQ ID: 387 (internal designation 105-073-2-0-A7-CS)

The protein of SEQ ID NO : 387 encoded by the cDNA of SEQ ID NO: 146 is expressed in liver, ovary, prostate and overexpressed in salivary glands. The protein of SEQ ID NO : 387 belongs to the abhydrolase family, and is caracterized by the alpha/beta hydrolase fold (Protein Eng 1992;5:197-211, which disclosure is hereby incorporated by reference in its entirety), that is common to a number of hydrolytic enzymes of widely differing phylogenetic origin and catalytic function.

The core of each enzyme is an alpha/beta-sheet (rather than a barrel), containing 8 strand connected by helices. The enzymes are believed to have diverged from a common ancestor, preserving the arrangement of the catalytic residues. All have a catalytic triad, the elements of which are borne on loops, which are the best conserved structural features of the fold.

Epoxide hydrolases are a family of enzymes which hydrolyze a variety of exogenous and endogenous epoxides to their corresponding diols. The epoxide hydrolase add water to epoxides, forming the corresponding diol. On the basis of sequence similarity, it has been proposed that the mammalian soluble epoxide hydrolase contain 2 evolutionarily distinct domains, the N-terminal domain is similar to bacterial haloacid dehalogenase, while the C-terminal domain is similar to soluble plant epoxyde hydrolase, microsomal epoxide hydrolase, and bacterial haloalcane dehalogenase (DNA Cell Biol. 14 :61-71 (1995), which disclosure is hereby incorporated by reference in its entirety. Human epoxide hydrolase catalyse the addition of water to epoxides to form the corresponding dihydrodiol. The enzymatic hydratation is essentially irreversible and produces mainly metabolites of lower reactivity that can be conjugated and excreted. The reaction of epoxide hydrolase is therefore generally regarded as detoxifying. Commonly the function of epoxide hydrolase is finally followed by excretion of the diols. However, reactivation of certain diols by a second epoxidation may happen. Epoxide hydrolase inactivates also the epoxides existing in the metabolism of endogenous compounds. Lipophilic xenobiotics tend to accumulate into tissues, and they must be transformed to water soluble compounds to enable the excretion. In this transformation process reactive intermediates are produced. If biotransformation fails to detoxify these reactive intermediates, they may react covalently with critical targets like the genetic material, or start harmful reaction chains like lipid peroxidation. Therefore, epoxide hydrolases are thought to be responsible for carcinogenicity and mutagenicity phenomenon (Exp Pathol 1990;39(3-4):195-6.). In addition, the interaction between epoxide hydrolase activity and alcohol-metabolizing enzymes, suggests that epoxide hydrolase activity may be associed with the susceptibility to alcoholic liver disease and hepatocellular carcinoma (Toxicol. Lett. 10 ;115 (1) :17-22 (2000), which disclosure is hereby incorporated by reference in its entirety). Compounds containing the epoxide functionality

have become common environmental contaminants because of their wide use as pesticides, sterilants, and industrial precursors. Such compounds also occur as products, by-products, or intermediates in normal metabolism and as the result of spontaneous oxidation of membrane lipids (i.e. see, Brash, et al., Proc. Natl. Acad. Sci., 85:3382-3386 (1988), and Sevanian, A., et al., Molecular Basis of Environmental Toxicology (Bhatnager, R. S., ed.) pp. 213-228, Ann Algor Science, Michigan (1980)). As three-membered cyclic ethers, epoxides are often very reactive and have been found to be cytotoxic, mutagenic and carcinogenic (i.e. see Sugiyama, S., et al., Life Sci. 40:225-231 (1987)). Cleavage of the ether bond in the presence of electrophiles often results in adduct formation. As a result, epoxides have been implicated as the proximate toxin or mutagen for a large number of xenobiotics. Reactions of detoxification using epoxide hydrolases typically decrease the hydrophobicity of a compound, resulting in a more polar and thereby excretable substance.

It is believed that the protein of SEQ ID NO: 387 or part thereof is an hydrolase, preferably an epoxyde hydrolase. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 387 from positions 2 to 132, 52 to 137, 29 to 120, 12 to 137, 19 to 136, 151 to 209, 141 to 209, 30 to 108, and 35 to 108. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 387 having any of the biological activity described herein. The hydrolytic activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in Cancer res 40(7):2552-6 (1980); Exp Pathol 39(3-4):195-6 (1990), which disclosures are hereby incorporated by reference in their entireties.

The invention also relates to methods and compositions using the protein of the invention or part thereof to diagnose, prevent and/or treat several disorders linked to overexpression of the protein of the invention including alcoholic liver disease, hepatocellular carcinoma, ovarian and prostate cancers.

In addition, the protein of the invention or part thereof may be used to identify inhibitors for mechanistic and clinical applications. Such inhibitors may then be used to identify or quantify the protein of the invention in a sample, and to diagnose, treat or prevent any of the disorders where the protein's hydrolytic activity is undesirable and/or deleterious such as disorders characterized by tissue degradation including but not limited to amyloidosis, colitis, lysosomal diseases, arthritis, muscular dystrophy, inflammation, tumor invasion, glomerulonephritis, parasite-borne infections, Alzheimer's disease, periodontal disease, and cancer metastasis.

In another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably ovarian, liver or prostate, more preferably salivary glands. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art. Such tissue specific antibodies may then be used to identify tissues of unknown origin, for example,

forensic samples, differentiated tumor tissue that metastasized to foreign bodily, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Protein of SEQ ID No: 398 (internal designation: 160-31-3-0-E4-CS)

The protein of SEQ ID No: 398 encoded by the cDNA of SEQ ID No: 157, is overexpressed in fetal brain and shows homology with diverse hydrolases. The protein of the invention also displays a motif characteristic of isochorismatase proteins from positions 17 to 147. In addition, the protein of the invention is an alternatively spliced form of an unnamed human protein.

It is believed that the protein of SEQ ID NO: 398 or part thereof is an hydrolase, preferably acting on ether bonds, more preferably an ether hydrolase. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 398 from positions 17 to 147. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 398 having any of the biological activity described herein. The hydrolytic activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in US patents 5,445,942; 5,445,956, 6,017,746 and 5,871,616 and in Rusnak et al, 1990; Biochemistry 29 1425-1435.

In another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably fetal brain. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Proteins of SEQ ID NOs: 260 and 265 (internal designation 116-004-3-0-A6-CS and 116-091-1-0-D9-CS respectively)

The protein of SEQ ID NO: 260 encoded by the cDNA SEQ ID NO: 19 and over expressed in liver and testis is an isoform of the protein of SEQ ID NO: 265 encoded by the cDNA SEQ ID NO: 24 over expressed in liver. Both proteins show homology to murine EPCS26 (Hemberger M. et al., Dev. Biol. 222, 158-169 (2000)) with Genbank accession number AF250838. The proteins of SEQ ID NO: 260 and 265 contain a signal peptide (cleavage site at position 18) that could allow the export of the protein to the extracellular domain, the export to a cellular membrane or to define a particular subcellular localization. The cDNA encoding EPCS26 has been shown to be differentially expressed during the process of trophoblast invasion.

Implantation and placentation are key processes in mammalian embryonic development. They physically connect the embryo to its mother and are critical for sufficient nutrient and gaz exchange. The extraembryonic cell lineage is the first to differentiate in the developing conceptus,

reflecting the importance of this cell for the establishment of fetal-maternal connections. During murine development, the outer layer of blastocyt, the mural trophectoderm, begins to differentiate into primary trophoblast giant cells on day 5 of gestation (e5). These cells invade the uterine epithelium and penetrate deeply into the stroma. At the same time, the polar trophectoderm cells continue to proliferate and form the ectoplacental cone. On e7, the outer cells of the ectoplacental cone begin to differentiate into secondary trophoblast giant cells. The invasion of uterine stroma by these cells is critical for successful placentation (Cross et al., Science 266, 1508-1518 (1994)).

Trophoblast invasion triggers secretion of proteinases that degrade extracellular matrix molecules. Mouse trophoblasts have been shown to synthesize and secrete serine proteases, matrix metalloproteinases and cysteine proteinases. Invasion of the trophoblast is a highly controlled process. The decidula restricts invasion by secreting proteinases inhibitors. Proteinases and proteinases inhibitors have antagonistic functions in implantation and placentation which may be mirrored by the reciprocity of their expression patterns (Alexander et al Development 122, 1723-1736 (1996)).

During tumor invasion and metastasis, the degradation of the basement membranes is often accomplished by the proteinases implicated in implantation and normal trophoblast invasion (Strickland and Richards Cell 71, 355-357 (1992), Wilson et al. Proc. Natl. Acad. Sci. USA 94, 1402-1407 (1997)). Uncontrolled trophoblast invasion, as in choriocarcinomas, results in one of the most metastatic tumors known (Strickland and Richards Cell 71, 355-357 (1992)).

A deficient fonction of the protein of the invention could result in an uncontrolled trophoblast invasion, and like in choriocarcinomas results in one of the most metastatic tumors known (Strickland and Richards Cell 71, 355-357 (1992)).

It is believed that the proteins of SEQ ID NO: 260 and 265 or part thereof play a role in proteolysis, preferably during embryogenesis, more preferably during trophoblast invasion. The proteins of the invention or part thereof may act as secreted proteinases that degrade extracellular matrix molecules or at the contrary as proteinase inhibitors. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 260 from positions 7 to 122 and the amino acids of SEQ ID NO: 265 from positions 7 to 81. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 260 and 265 having any of the biological activities described herein. The proteolytic activity of the proteins of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in US patent 6,069,229 and 5,861,267. The protease inhibitor activity of the proteins of the invention or part thereof may be assayed using any of the assays known to those skilled in the art and using methods for determining inhibition constants well known to those skilled in the art (see Fersht, ENZYME STRUCTURE AND MECHANISM, 2nd ed., W.H. Freeman and Co., New York, (1985))

In addition, the proteins of the invention or part thereof may be used to diagnose, treat or prevent any of the disorders characterized by undesirable and/or deleterious hydrolytic activity such as disorders characterized by tissue degradation including but not limited to amyloidosis, colitis, lysosomal diseases, arthritis, muscular dystrophy, inflammation, tumor invasion, glomerulonephritis, parasite-borne infections, Alzheimer's disease, periodontal disease, cancer metastasis, and choriocarcinoma. For diagnostic purposes, the expression of the proteins of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. Alternatively, inhibitors for the proteins' activity may be developed and use to inhibit and/or reduce its activity using any methods known to those skilled in the art. Overexpression of the proteins of the invention or part thereof may be achieved using any of the gene therapy method described herein.

In another embodiment, the invention relates to methods and compositions using the protein of the inventions or part thereof as a marker protein to selectively identify tissues, preferably liver and testis for the protein of SEQ ID NO: 260, preferably liver for the protein of SEQ ID NO: 265. For example, the proteins of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Protein of SEQ ID NO: 265 (internal designation 116-088-4-0-A9-CS)

The protein of SEQ ID NO: 265 encoded by the cDNA of SEQ ID NO: 24 is overexpressed in testis and liver. This protein of the invention is homologous to the GdX protein, also named UBL4 (Toniolo et al., Proc Natl Acad Sci USA 1988;85:851-5), found in both human (GENPEPT accession number L44140) and mice species (GENPEPT accession number J04761). In addition, the 174-amino-acid-long protein of SEQ ID NO: 265, which is similar in size to ubiquitin-like proteins, displays a pfam consensus domain from position 1 to 82 that is the hallmarks of ubiquitin family proteins.

Ubiquitin is a protein of 76 amino acid residues, found in all eukaryotic cells, and which is extremely well conserved from protozoan to vertebrates (Jentsch et al. Trends Cell Biol 2000;10:335-42). It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response, ribosome biogenesis, cell-cycle progression, signal transduction, transcription and antigen presentation (Wilkinson et al. Annu Rev Nutr 1995;15:161-89). The first ubiquitin is covalently ligated to target proteins through an isopeptide linkage between the C-terminal glycine residue of ubiquitin and an internal ε-amino group of lysine residue of the substrate. To generated an efficient proteasomal targeting signal, additional ubiquitin are linked to

the first one by isopeptide bounds, and form branched poly-ubiquitin complexes (Thrower et al. EMBO J 2000;19: 94-102). Covalent binding of ubiquitin to proteins marks them for subsequent degradation by a multicomponent enzymatic complex known as the 26S proteasome (Hershko et al. Annu Rev Biochem 1992;61:761-807).

The genes coding ubiquitin-like proteins fall into two separate classes (Hershko et al. Annu Rev Biochem 1992;61:761-807). Proteins of the first class are frequently designed as ubiquitin-like modifiers, or UBLs. They produce polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin, with a variable number of repeats. These linear polymer of ubiquitin are linked covalently through peptide bonds between the C-terminal glycine residue and N-terminal lysine residue of contiguous ubiquitin molecules. Proteins of the second class are habitually named as ubiquitin-domain proteins, or UDPs. These proteins bear a single domain of the N-terminal domain that is related to ubiquitin, fused to a C-terminal ribosomal domain consisting of 52 or 76-80 amino-acid residues (Finley et al. Nature 1989;338:394-401). These proteins are not conjugated to other proteins and function as an heterogeneous group of proteins. To date, this family includes RAD23, DSK2, PLIC-1, PLIC-2/Chap1, XDRP1, BAG-1, BAT3/Chap2, Scythe, Parkin, UIP28, UBP6, Elongin B, and GdX. In addition, the protein of invention of SEQ ID NO: 265 clearly belongs to the UDPs family, as it displays a single ubiquitin N-terminal consensus domain, which is the hallmark of this protein family subset.

UDPs participate to regulation of proteolysis through multiple mechanisms such as interaction with catalytically active 26S proteasome for RAD23 (Schauber et al. Nature 1998;391:715-8), hPLIC-1 and hPLIC-2 (Kleijnen et al. Mol Cell 2000;6:409-19), and BAG-1 (Luders et al. J Biol Chem 2000;275:4613-7), removing ubiquitin from conjugates for UBP6 (Wyndham et al. Protein Sci 1997;8:1268-75) and negative regulation of multi-ubiquitin chain assembly for RAD23 (Ortolan et al. Nature cell Biol 2000; 2:601-8). In addition, an increasing body of evidence indicates that some UDPs participate to other cellular functions as protein folding (Luders et al. J Biol Chem 2000;275:4613-7), apoptosis (Kaye et al. FEBS Lett 2000;467:348-55), and nucleotide-excision repair (de Laat et al. Genes Dev 1999;13:768-785). UDPs family proteins have been shown directly associated with pathogenesis of several diseases including xeroderma pigmentosum for RAD23 (Masutani et al. EMBO J 1994;13:1831-43), and Parkinson's disease for parkin (Kitada et al. Nature 1998;392:605-8). In addition, involvement of ubiquitin-like proteins or abnormal ubiquitinated accumulation of proteins has been found in multiple human disorders. Most of them, but not all, involve nervous central system as Alzheimer's disease (van Leeuwen et al. Science 1998;279:242-7), diffuse Lewy body disease (Iseki et al. J Neurol Sci 1997;146:53-7), Huntington disease (Scherzinger et al. Cell 1997;90:549-58), and amyotrophic lateral sclerosis (Leigh et al. Brain 1991;114:775-88). In most disorders, ubiquinated-proteins accumulate within cells and form aggregates termed inclusion bodies that have characteristic appearance on histological examination. In addition, abnormal accumulation of ubiquitinated proteins has been

found in Von-Hippel Lindau disease (Kamura et al. Proc Natl Acad Sci USA. 2000;97:10430-5), and in liver of alcoholic hepatitis patients (Ohta et al. Lab Invest. 1988;59:848-56). Components of hepatocytes are released within the circulation in alcoholic hepatitis (Sorbi et al. Am J Gastroenterol 1999;94:1018-22)

It is believed that the protein of SEQ ID NO: 265 or part thereof plays a role in the regulation of proteolysis, preferably as a ubiquitin-like protein, more preferably as a ubiquitin-domain protein. In addition, the protein of the invention may play a role in protein folding, apoptosis and nucleotide-excision repair. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 265 from positions 1 to 82. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 265 having any of the biological activity described herein.

In an embodiment, the invention relates to compositions and methods using the protein of the invention or part thereof to remove, identify or inhibit contaminating proteases in a sample. Compositions comprising the polypeptides of the present invention may be added to biological samples as a "cocktail" with other protease inhibitors to prevent degradation of protein samples. The advantage of using a cocktail of protease inhibitors is that one is able to inhibit a wide range of proteases without knowing the specificity of any of the proteases. Using a cocktail of protease inhibitors also protects a protein sample from a wide range of future unknown proteases which may contaminate a protein sample from a vast number of sources. Such protease inhibitor cocktails (see for example the ready to use cocktails sold by Sigma) are widely used in research laboratory assays to inhibit proteases susceptible of degrading a protein of interest for which the assay is to be performed. For example, the protein of the invention or part thereof is added to samples where proteolytic degradation by contaminating proteases is undesirable. Alternatively, the protein of the invention or part thereof may be bound to a chromatographic support, either alone or in combination with other protease inhibitors, using techniques well known in the art, to form an affinity chromatography column. A sample containing the undesirable protease is run through the column to remove the protease. Alternatively, the same methods may be used to identify new proteases.

Another embodiment of the invention relates to compositions and methods of using the protein of invention or part thereof to develop assays for the immunohistochemical detection of testicular malignant tissue, as the protein is overexpressed in such tissue. For instance, this could be used for staging lymph node testicular cancer dissemination using the techniques and methods detailed in Nazeer et al. Oncol Rep (1998);5:1425-9. The ability to specifically visualize malignant tissues (and cells derived from the tissues), is useful for numerous applications, including to determine the origin, to identity e.g. cancerous cells, as well as to facilitate the identification of particular cells and tissues for, e.g. the evaluation of histological slides.

In another embodiment, the invention relates to compositions or methods using the protein of SEQ ID NO: 265 or part thereof to diagnose, treat and/or prevent disorders including, but not limited to xeroderma pigmentosum, Von-Hippel Lindau disease, alcoholic hepatitis, in neurodegenrative diseases such as Alzheimer's disease, diffuse Lewy body disease, Huntington disease, and amyotrophic lateral sclerosis. Detection of poly-ubiquinated protein conjugates in biological samples, such as brain tissues for the diagnosis of neurodegenerative disorders or liver and serum or plasma for the diagnosis of alcoholic hepatitis, may be performed using antibodies or nucleic acid able to detect the expression of the protein of the invention using immunohistochemisty, enzyme-linked immunosorbant assay (ELISA) or any other technique known to those skilled in the art including Northern blotting, RT-PCR or immunoblotting methods described herein as well as the technique described in Mimnaugh et al. Electrophoresis 1999;20:418-28. The expression of the protein of the invention in patients' samples is then compared to the expression in control individuals.

In still another embodiment, the invention relates to compositions or methods to treat, attenuate and/or prevent disorders including, but not limited to xeroderma pigmentosum, Von-Hippel Lindau disease, alcoholic hepatitis, in neurodegenerative diseases such as Alzheimer's disease, diffuse Lewy body disease, Huntington disease, and amyotrophic lateral sclerosis using the protein of the invention, part thereof, or any other compounds developed using the present protein as nucleic acids, antibodies, or chemical substances. In a preferred embodiment, proteins or other compounds targeted against the protein of invention or part thereof may be used to treat, prevent and/or attenuate disorders in which ubiquitin-like proteins or abnormal accumulation of ubquitininated proteins has been found and can be involved in pathogenesis of the disease. For instance, proteins or other compounds targeted against protein of SEQ ID NO: 265 can be administered to treat or attenuate symptoms of patients affected with Alzheimer's disorder or any other neurodegenerative disorders.

Protein of SEQ ID NO: 408 (internal designation 174-8-2-0-C10-CS)

The protein of SEQ ID NO: 408 encoded by the cDNA of SEQ ID NO: 167 found in salivary gland and brain is homologous to a drosophila melanogaster protein thought to be transmembraneous (STR: Q9V641). The 345-amino-acid-long protein of SEQ ID NO: 408 displays the Rhomboid pfam domain from positions 186 to 323 and is predicted having six transmembrane domains from positions 101 to 121, 167 to 187, 204 to 224, 243 to 263, 273 to 293, 298 to 318.

Rhomboid genes were identified in flies and in organisms as diverse as Arabidopsis, yeast, bacteria, and mammals. Human and rat homologues of Rhomboid have been identified (Pascal et al.: 1998; FEBBS Lett. 429; 337-340). This very widespread conservation implies that the Rhomboid family proteins have a fundamental function within many cells. The Drosophila

Rhomboid has six transmembrane domains and an amino terminal hydrophobic region like the protein of the invention.

The 355-amino-acid-long Drosophila Rhomboid protein is known to control many aspects of fly development and especially, to establish position along the dorsoventral axis and then again later to specify the fate of neuronal precursor cells. Rhomboid expression is sufficient to activate EGF receptor (EGFr) signaling in all tissues in Drosophila, while loss of Rhomboid mimics reduction (or loss) of EGFr signaling in almost all tissues (Guichard et al :.1999 Development; 126, 2663-2676). As in mammals, the drosophila EGF receptor controls many aspects of growth and development. Three activating ligands of the drosophila EGFr have been described, the most developmentally significant being the TGF alpha-like molecule, Spitz (Rutledge et al. : 1992; Genes & Dev. 6; 1503-1517). None of the Rhomboid-like proteins from species other than Drosophila have clearly assigned functions. However, there is compelling genetic evidence from Drosophila that Rhomboid has a key role in intercellular signaling: it functions as an activator of the EGF receptor, probably by controlling the activation the TGF-like ligand Spitz (Guichard et al :.1999 Development; 126, 2663-2676). Indeed, Rhomboid expression is the principal rate-limiting step in activation of the Ras/MAP kinase pathway by the EGFr.

Like mammalian TGF alpha, Spitz is synthesized as a functionally inert transmembrane protein; subsequently, the proteolytic release of the extracellular portion of the molecule gives rise to a soluble and potent EGFr ligand (Golembo et al.: 1996; Development; 122; 3363-3370). Unlike all other essential components of EGFr signaling, the expression of Rhomboid is tightly restricted to sites of signaling activity. It has been proposed that Rhomboid attains its key role in the pathway by regulating the proteolytic cleavage of Spitz (Wasserman et al : 2000; Genes & development; 14; 1651-1663). The preeminence of Rhomboid in a pathway as critical to development and growth control as the EGFr/Ras/Map kinase cascade provides a strong incentive to understand its molecular mechanism. By analogy to mammalian EGFr ligands that are similarly processed, Spitz cleavage is expected to be catalyzed by an ADAM like protease (Black et al 1998; Curr. Opin. Cell.Biol. 10; 654-659), but Rhomboid resembles no known protease.

The Drosophila eye has served as a useful model for studying mechanisms of EGFr and Ras signaling. At least five different roles for the receptor have been identified (for reviews see Wasserman et al : 2000; Genes & development; 14; 1651-1663), the best characterized being its function in recruiting cells into the developing ommatidium- the individual unit of the fly compound eye. Each ommatidium contains eight photoreceptors, four cone cells that secrete lens material, and an average of eight pigment cells. It has been shown that the fly EGFr has a role in regulating cell survival in the developing eye (Dominguez et al. 1998; Curr. Biol. 8; 1039-1048).

The EGFr signaling pathway has been conserved between flies and vertebrates. The EGFr family consists of four members, HER1 (c-erbB1, EGFR), HER2 (c-erbB2), HER3 (c-erbB3), HER4 (c-erbB4), expressed in a wide range of cells (Gullick W.J. 1998; Br. Cancer Res. Treat.; 52,

43-53). TGF.alpha. and its homologs have been found to be the most abundant ligands for the EGF/TGF.alpha. receptor in most parts of the brain (Kaser, et al., (1992)Brain Res Mol Brain Res: 16:316-322). There appears to be a widespread distribution of TGF.alpha. in various regions of the brain in contrast to EGF which is only present in smaller, more discrete areas, suggesting that TGF-alpha might play a physiological role in brain tissues. These numerous receptor sites for TGF.alpha in the brain suggest that TGF has an important utility in promoting normal brain cell differentiation and function.

Transforming growth factor alpha (TGF-alpha.) is a relative of epidermal growth factor (EGF) and like EGF, it exerts its effects on cells through binding to the EGF receptor. The precise physiological roll of TGF.alpha. is still not clear, although it appears to be important in eye and hair follicle development and may play a role in both the immune system and in wound healing. (See Kumar, et al.; 1995 Cell Biology International, 19:5, 373-388). The EGF family receptors currently includes four EGF receptors. The EGFR2 receptor may also be referred to as ERB-2 and this molecule is useful for a variety of diagnostic and therapeutic indications (Prigent, S. A., and Lemoine, N. R., (1992) Prog Growth Factor Res., 4:1-24). The TGF-alpha is likely a ligand for one or more of these receptors as well as for yet an identified new EGF-type receptor. Use of the TGF-alpha. can assist with the identification, characterization and cloning of such receptors. For example, the EGF receptor gene represents the cellular homolog of the v-erb-B oncogene of avian erythroblastosis virus. Over expression of the EGF-receptor or deletion of kinase regulatory segments of the protein can bring about tumorigenic transformation of cells (Manjusri, D. et al., (1991) Human Cytokines, 364 and 381).

The EGF receptor, and the related ErbB family of receptor tyrosine kinases, have indeed been much implicated in human cancer. It is commonly believed that hyperactive receptor signaling promotes dysregulates growth control and in involved in the onset of malignancy, as well as in the disruption of developmental programs. Very little, however, is known about ErbB physiological regulation in humans. The fruitfly, Drosophila melanogaster, has a single receptor homologous to the four ErbB receptors. As signaling mechanisms have been well conserved between flies and mammals, these results of experiments in flies are relevant to the study of the human receptors in development and disease. Two areas of recent progress are emphasized. First, a number of signal modulators have been identified, including three EGF receptor inhibitors, several of which have human homologues. Second, the signaling molecules are integrated into regulatory networks that specify the elaborate activation profiles needed in development (positive and negative feedback control of EGF receptor signaling emerges as a central theme).

It is thus important to discover whether Rhomboid-like proteins also have functions similar to those observed in Drosophila in other higher organisms, including mammals, because of the substantial clinical importance of the EGFr pathway.

It is believed that the protein of SEQ ID NO: 408 or part thereof plays a role into the control of cellular signaling. Preferably the protein of the invention or part thereof plays a role in the activation of EGFr-mediated cell signaling, probably through the control of the activation of EGFr ligands, such as EGF, TGF alpha and TGF alpha-like factor, more probably through the proteolytic cleavage of such ligands. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 408 from positions 186 to 323, 101 to 121, 167 to 187, 204 to 224, 243 to 263, 273 to 293, and 298 to 318. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 408 having any of the biological activity described herein. The proteolytic activity of the protein of the invention or part thereof as well as its involvment in regulation of cellular signalling though the activation of EGFr may be assayed using any of the assays known to those skilled in the art.

An embodiment of the invention relates to composition and methods using the protein of the invention or part thereof to identify and/or quantify the activation of EGF receptors, preferably veretebrate EGF receptors, more preferably human ErbB receptors, in a biological sample, and thus used in assays and diagnostic kits for the quantification of such activation in bodily fluids, in tissue samples, and in mammalian cell cultures. The assessment of the activation of EGF receptors may be perfomed using any assay familiar to those skilled in the art. Preferably, a defined quantity of the protein of the invention or part thereof is added to the sample under conditions allowing the activation of EGFr. Then, the activation of EGFr is assayed and eventually compared to a control using any of the techniques known by those skilled in the art.

The present invention also relates to diagnostic assays for detecting altered levels of the protein of the present invention in various tissues since an over-expression of the proteins compared to normal control tissue samples can detect the presence of certain disease conditions such as neoplasia, skin disorders, ocular disorders and inflammation. Assays used to detect levels of the polypeptide of the present invention in a sample derived from a host are well-known to those of skill in the art and include radioimmunoassays competitive-binding assays, Western Blot analysis and preferably ELISA assays.

This invention is also related to the use of SEQ ID No: 167 or its complement as a diagnostic tool. Detection of a mutated form of the nucleotide sequence of SEQ ID No: 167 of the present invention will allow a diagnosis of a disease or a susceptibility to a disease which results from underexpression of the polypeptide of the present invention for example, improper wound healing, improper neurological functioning, ocular disorders, kidney and liver disorders, hair follicular development, angiogenesis and embryogenesis Individuals carrying mutations in the human nucleotide sequence of SEQ ID No: 167 of the present invention may be detected at the DNA level by a variety of techniques. Nucleic acids for diagnosis may be obtained from a patient's cells, such as from blood, urine, saliva, tissue biopsy and autopsy material. The genomic DNA may be used directly for detection or may be amplified enzymatically by using PCR (Saiki et al., (1986)

Nature, 324:163-166) prior to analysis. RNA or cDNA may also be used for the same purpose. As an example, PCR primers complementary to the nucleic acid encoding a polypeptide of the present invention can be used to identify and analyze mutations thereof. For example, deletions and insertions can be detected by a change in size of the amplified product in comparison to the normal genotype. Point mutations can be identified by hybridizing amplified DNA to radiolabeled RNA or alternatively, radiolabeled antisense DNA sequences. Perfectly matched sequences can be distinguished from mismatched duplexes by RNase A digestion or by differences in melting temperatures.

In another embodiment, the protein of the invention or part thereof can be used to diagnose, treat and/or prevent disorders linked to dysregulation of growth control, such as cancer and other disorders relating to abnormal cellular differentiation, proliferation, or degeneration, including hyperaldosteronism, hypocortisolism (Addison's disease), hyperthyroidism (Grave's disease), hypothyroidism, colorectal polyps, gastritis, gastric and duodenal ulcers, ulcerative colitis, and Crohn's disease, neurodegenrative disroders such as Parkinson's and Alzheimer's diseases using any methods and/or techniques described herein. For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. In addition, the protein of the invention or part thereof may be used to evaluate the disease progression and the clinical treatment efficiency. Inhibition of expression of the protein of the invetion or part thereof to inhibit EGFr activation could be achieved by many means known to those skilled in the art including those described in the present application such as antisense nucleotide or triple helix strategies.

Protein of SEQ ID NO:291 (internal designation: 180-19-4-0-F4-CS)

The protein of SEQ ID No:291 encoded by the cDNA of SEQ ID No:50 is homologous to proteins of the tissue inhibitor of metalloproteinases (TIMP) family. The protein of the invention (207 amino-acids) is highly homologous to and appears to be a variant of the metalloproteinase inhibitor 1 precursor (TIMP-1, 207 amino-acids) human protein (SwissProt P01033). The protein of the invention is stronly expressed in the liver, ovary and testis.

There are many different types of collagen found in the body and they, together with other extracellular matrix components, such as elastin, gelatin, proteoglycan and fibronectin, make up a large proportion of the body's extracellular tissue. Matrix metalloproteinases (MMPs) are enzymes that are involved in the degradation and denaturation of extracellular matrix components. Collagenases, for example, are MMPs that degrade or denature collagen. A large number of different collagenases are known to exist. These include interstitial collagenases, type IV-specific collagenases and collagenolytic proteinases. Collagenases are generally specific for collagens which, in their full triple helix structure, are extremely resistant to other enzymes. Other MMPs are

involved in the degradation and denaturation of different extracellular matrix components, for example, elastin, gelatin and proteoglycan. Some MMPs are able to degrade or denature several different types of collagen and also other extracellular matrix components. For example, stromelysin degrades type IV collagen, which is found in basement membrane, and also has an effect on other extracellular matrix components such as elastin, fibronectin and cartilage proteoglycans. The ability of MMPs metalloproteinases (such as collagenase, stromelysin, and gelatinase) to degrade various components of connective tissue makes them potential targets for controlling numerous pathological processes.

The presence of tissue inhibitors of MMPs has been observed in a variety of explants and in monolayer cultures of mammalian connective tissue cells (Vater et al 1979 and Stricklin and Wegus 1983). Not only collagenase inhibitors but also inhibitors for other MMPs, for example, gelatinase and proteoglycanase have been found. MMP inhibitors are generally unable to bind the inactive (zymogen) forms of the respective enzymes but complex readily with active forms (Murphy et al 1981). Tissue MMP inhibitors are found, for example, in dermal fibroblasts, human lung, gingival, tendon and corneal fibroblasts, human osteoblasts, uterine smooth muscle cells, alveolar macrophages, amniotic fluid, plasma, serum and the .alpha.-granule of human platelets (Stricklin and Wegus 1983; Welgus et al 1985; Welgus and Stricklin 1983; Bar-Sharvit et al 1985; Wooley et al; 1976; and Cooper et al 1985).

The protein of the invention is a secreted TIMP-1 protein which tightly complexes with metalloproteinases and irreversibly inactivate them. TIMP-1 has been identified as a secretory product of platelets and alveolar macrophages

Thus, an embodiment of the present invention relates to the use of the protein of the invention or a fragment thereof to inhibit the action of MMPs by directly inhibiting the enzyme activity like a conventional inhibitor. The inhibitory activity of a MMP inhibitor may be assessed by any method suitable for determining inhibitory activity of a compound with respect to an enzyme. Such methods are described in standard textbooks of biochemistry.

In one embodiment, the protein of SEQ ID NO:291 can be used to treat and diagnose disorders associated with excessive MMP expression, such as inflammatory disorders such as rheumatoid arthritis, osteoarthritis, osteopenias such as osteoporosis, pulmonary emphysema, periodontitis, gingivitis, corneal epidermal or gastric ulceration, and tumour metastasis, invasion and growth, Paget's disease, hyperparathyroidism. MMP inhibitors are also of potential value in the treatment of neuroinflammatory disorders, including those involving myelin degradation, for example multiple sclerosis, as well as in the management of angiogenesis dependent diseases, which include arthritic conditions and solid tumour growth as well as psoriasis, proliferative retinopathies, neovascular glaucoma, ocular tumours, angiofibromas and hemangiomas. The present invention relates to a method of treating diseases in which MMPs are involved such as atherosclerotic plaque rupture, restenosis, aortic aneurysm (including abdominal aortic aneurysm

and brain aortic aneurysm), congestive heart failure, left ventricular dilatation, myocardial infarction, decubital ulcers, chronic ulcers or wounds, renal disease, or other autoimmune or inflammatory diseases dependent upon tissue invasion by leukocytes, Crohn's disease, acute respiratory distress syndrome, asthma, chronic obstructive pulmonary disease, Alzheimer's disease, organ transplant toxicity, cachexia, allergic reactions, allergic contact hypersensitivity, epidermolysis bullosa, loosening of artificial joint implants, stroke, cerebral ischemia, head trauma, spinal cord injury, neuro-degenerative disorders (acute and chronic), Huntington's disease, Parkinson's disease, migraine, depression, peripheral neuropathy, pain, cerebral amyloid angiopathy, nootropic or cognition enhancement, amyotrophic lateral sclerosis, ocular angiogenesis, macular degeneration, abnormal wound healing, burns, diabetes, scleritis, AIDS, sepsis, septic shock.

In another embodiment, the protein of SEQ ID NO:291 has potential value in the treatment or diagnosis of atherosclerosis. The rupture of atherosclerotic plaques is the most common event initiating coronary thrombosis. Destabilization and degradation of the extracellular matrix surrounding these plaques by MMPs has been proposed as a cause of plaque fissuring. The shoulders and regions of foam cell accumulation in human atherosclerotic plaques show locally increased expression of gelatinase B, stromelysin-1, and interstitial collagenase. In situ zymography of this tissue revealed increased gelatinolytic and caseinolytic activity (Galla, et al., J. Clin. Invest., 1994;94:2494-2503). In addition, high levels of stromelysin RNA message have been found to be localized to individual cells in atherosclerotic plaques removed from heart transplant patients at the time of surgery (Henney, et al., Proc. Nat'l. Acad. Sci., 1991;88:8154-8158).

In another embodiment, the protein of the invention has utility in treating or detecting degenerative aortic disease associated with thinning of the medial aortic wall. Increased levels of the proteolytic activities of MMPs have been identified in patients with aortic aneurysms and aortic stenosis (Vine N. and Powell J. T., Clin. Sci., 1991;81:233-239).

In another embodiment, the protein of the invention can be used as a treatment or diagnostic tool for heart failure and associated ventricular dilatation. Heart failure arises from a variety of diverse etiologies, but a common characteristic is cardiac dilation which has been identified as an independent risk factor for mortality (Lee, et al., Am. J. Cardiol., 1993;72:672-676). This remodeling of the failing heart appears to involve the breakdown of extracellular matrix. MMPs are increased in patients with both idiopathic and ischemic heart failure (Reddy, et al., Clin. Res., 1993;41:660A; Tyagi S. C., et al., Clin. Res., 1993;41:681A). Animal models of heart failure have shown that the induction of gelatinase is important in cardiac dilation (Armstrong, et al., Can. J. Cardiol., 1994;10:214-220), and cardiac dilation precedes profound deficits in cardiac function (Sabbah, et al., Am. J. Physiol., 1992;263:H266-H270).

In another embodiment, the protein of the invention is useful in treating or detecting neointimal proliferation, leading to restenosis, frequently developed after coronary angioplasty.

The migration of vascular smooth muscle cells (VSMCs) from the tunica media to the neointima is a key event in the development and progression of many vascular diseases and a highly predictable consequence of mechanical injury to the blood vessel (Bendeck M. P., et al., Circulation Research, 1994;75:539-545). Northern blotting and zymographic analyses indicated that gelatinase A was the principal MMP expressed and excreted by these cells. Further, antisera capable of selectively neutralizing gelatinase A activity also inhibited VSMC migration across basement membrane barrier. (Pauly R. R., et al., Circulation Research, 1994;75:41-54).

In another embodiment, the protein of the invention is used to ensure normal kidney function, which is dependent on the maintenance of tissues constructed from differentiated and highly specialized renal cells. Those cells are in a dynamic balance with their surrounding extracellular matrix (ECM) components (Davies M. et al., Kidney Int., 1992;41:671-678). Effective glomerular filtration requires that a semi-permeable glomerular basement membrane (GBM) composed of collagens, fibronectin, enactin, laminin and proteoglycans is maintained. A structural equilibrium is achieved by balancing the continued deposition of ECM proteins with their degradation by specific MMPs. These proteins are first secreted as proenzymes and are subsequently activated in the extracellular space. These proteinases are in turn subject to counter balancing regulation of their activity by naturally occurring inhibitors as TIMPs.

Deficiency or defects in any component of the filtration barrier may have catastrophic consequences for longer term renal function. For example, in hereditary nephritis of Alport's type, associated with mutations in genes encoding ECM proteins, defects in collagen assembly lead to progressive renal failure associated with splitting of the GBM and eventual glomerular and interstitial fibrosis. In contrast, in inflammatory renal diseases such as glomerulonephritis, cellular proliferation of components of the glomerulus often precede obvious ultrastructural alteration of the ECM matrix. Cytokines and growth factors implicated in proliferative glomerulonephritis such as interleukin-1, tumor necrosis factor, and transforming growth factor beta can upregulate metalloproteinase expression in renal mesangial cells (Martin J. et al., J. Immunol., 1986;137:525-529; Marti H. P. et al., Biochem. J., 1993;291:441-446; Marti H. P. et al., Am. J. Pathol., 1994;144:82-94). These metalloproteinases are believed to be intimately involved in the aberrant tissue remodeling and cell proliferation characteristic of renal diseases, such as, IgA nephropathy which can progress to through a process of gradual glomerular fibrosis and loss of functional GBM to end-stage renal disease. Metalloproteinase expression has already been well-characterized in experimental immune complex-mediated glomerulonephritis such as the anti-Thy 1.1 rat model (Bagchus W. M., et al., Lab. Invest., 1986;55:680-687; Lovett D. H., et al., Am. J. Pathol., 1992;141:85-98).

In another embodiment, the protein of the invention can be used as a treatment or diagnostic tool for gingiva. Collagenase and stromelysin activities have been demonstrated in fibroblasts isolated from inflamed gingiva (Uitto V. J., et al., J. Periodontal Res., 1981;16:417-424), and

enzyme levels have been correlated to the severity of gum disease (Overall C. M., et al., J. Periodontal Res., 1987;22:81-88).

In another embodiment, the protein of the invention is useful for treating or detecting ulcers. Proteolytic degradation of extracellular matrix has been observed in corneal ulceration following alkali burns (Brown S. I., et al., Arch. Opthalmol., 1969;81:370-373). Thiol-containing peptides inhibit the collagenase isolated from alkali-burned rabbit corneas (Burns F. R., et al., Invest. Opththamol., 1989;30:1569-1575). Stromelysin, a member of the MMP family, is produced by basal keratinocytes in a variety of chronic ulcers (Saarialho-Kere U. K., et al., J. Clin. Invest., 1994;94:79-88). Stromelysin-1 mRNA and protein were detected in basal keratinocytes adjacent to but distal from the wound edge in what probably represents the sites of proliferating epidermis. Stromelysin-1 may thus prevent the epidermis from healing.

In another embodiment, the protein of the invention can be used as a treatment or diagnostic tool for tumor angiogenesis. Inhibitors of MMPs have shown activity in models of tumor angiogenesis (Taraboletti G., et al., Journal of the National Cancer Institute, 1995;87:293; and Benelli R., et al., Oncology Research, 1994;6:251-257). Davies et al., (Cancer Res., 1993;53:2087-2091) reported that a peptide decreased the tumor burden and prolonged the survival of mice bearing human ovarian carcinoma xenografts. A peptide of the conserved MMP propeptide sequence was a weak inhibitor of gelatinase A and inhibited human tumor cell invasion through a layer of reconstituted basement membrane (Melchiori A., et al., Cancer Res., 1992;52:2353-2356), and the natural tissue inhibitor of metalloproteinase-2 (TIMP-2) also showed blockage of tumor cell invasion in in vitro models (DeClerck Y. A., et al., Cancer Res., 1992;52:701-708). Studies of human cancers have shown that gelatinase A is activated on the invasive tumor cell surface (Strongin A. Y., et al., J. Biol. Chem., 1993;268:14033-14039) and is retained there through interaction with a receptor-like molecule (Monsky W. L, et al., Cancer Res., 1993;53:3159-3164).

In another embodiment, the protein of the invention can be used to treat and diagnose rheumatoid arthritis. Collagenases have been implicated in a number of diseases, including, rheumatoid arthritis (Mullins, D. E. et al 1983), and it has been proposed to use MMP inhibitors in the treatment of this condition. Several investigators have demonstrated consistent elevation of stromelysin and collagenase in synovial fluids from rheumatoid and osteoarthritis patients as compared to controls (Walakovits L. A., et al., Arthritis Rheum., 1992;35:35-42; Zafarullah M., et al., J. Rheumatol., 1993;20:693-697). TIMP-1 and TIMP-2 prevented the formation of collagen fragments, but not proteoglycan fragments, from the degradation of both the bovine nasal and pig articular cartilage models for arthritis, while a synthetic peptide hydroxamate could prevent the formation of both fragments (Andrews H. J., et al., Biochem. Biophys. Res. Commun., 1994;201:94-101).

In another embodiment, the protein of the invention is used to treat or diagnose inflammation. Gijbels et al., (J. Clin. Invest., 1994;94:2177-2182) recently described a peptide that

suppressed the development or reversed the clinical expression of experimental allergic encephalomyelitis (EAE) in a dose dependent manner, suggesting the use of MMP inhibitors in the treatment of autoimmune inflammatory disorders such as multiple sclerosis. A recent study by Madri has elucidated the role of gelatinase A in the extravasation of T-cells from the blood stream during inflammation (Ramanic A. M. and Madri J. A., J. Cell Biology, 1994;125:1165-1178). This transmigration past the endothelial cell layer is coordinated with the induction of gelatinase A and is mediated by binding to the vascular cell adhesion molecule-1 (VCAM-1). Once the barrier is compromised, edema and inflammation are produced in the CNS. Leukocytic migration across the blood-brain barrier is known to be associated with the inflammatory response in EAE. Inhibition of the metalloproteinase gelatinase A would block the degradation of extracellular matrix by activated T-cells that is necessary for CNS penetration. These studies provided the basis for the belief that an inhibitor of stromelysin-1 and/or gelatinase A will treat diseases involving disruption of extracellular matrix resulting in inflammation due to lymphocytic infiltration, inappropriate migration of metastatic or activated cells, or loss of structural integrity necessary for organ function.

The present invention provides the use of an MMP inhibitor in the manufacture of a medicament for the treatment or prophylaxis of scars. Collagen is the major component of scar and other contracted tissue and as such is the most important structural component to consider. Contraction of tissues comprising extracellular matrix components, especially of collagen-comprising tissues, may occur in connection with many different pathological conditions and with surgical or cosmetic procedures. Contracture, for example, of scars, may cause physical problems, which may lead to the need for medical treatment, or it may cause problems of a purely cosmetic nature.

During experiments on in vitro models of scar contraction, collagen appears to be invaded and permanently remodelled by fibroblasts and that such invasion and remodelling is inhibited by collagenase inhibitors. The remodelling generally appears as contraction of the collagen, the contraction of which is inhibited by inhibition of collagenase. Furthermore, inhibition of other MMPs also results in inhibition of contraction.

The present invention also provides the use of an MMP inhibitor in the treatment or prophylaxis of a natural or artificial tissue comprising extracellular matrix components to inhibit, i.e. restrict, hinder or prevent, contraction of the tissue, especially contraction resulting from a pathological condition or from surgical or cosmetic treatment.

Cosmetic treatments, such as chemical or physical dermal abrasion, used as anti-ageing treatments, cause trauma to the skin. Use of MMP inhibitors during the healing process which occurs after the initial abrasion is a cosmetic use of MMP inhibitors according to the present invention.

The present invention also provides the use of an MMP inhibitor to inhibit, i.e. restrict, hinder or prevent, invasion by cells, especially fibroblasts, into tissue comprising an extracellular

matrix and/or migration by cells, especially fibroblasts, in or through tissue comprising an extracellular matrix.

In another embodiment, the present protein is used to prevent or reduce contracture of scar tissue resulting from eye surgery. Glaucoma surgery to create new drainage channels often fails due to scarring and contraction of tissues. A method of preventing contraction of scar tissue formed in the eye, such as the application of a suitable agent, is therefore invaluable. Such an agent may also be used in the control of the contraction of scar tissue formed after corneal trauma or corneal surgery, for example laser or surgical treatment for myopia or refractive error in which contraction of tissues may lead to inaccurate results. It is also useful in cases where scar tissue is formed on/in the vitreous humor or the retina, for example, that which eventually causes blindness in some diabetics and that which is formed after detachment surgery, called proliferative vitreoretinopathy. Other uses include where scar tissue is formed in the orbit or on eye and eyelid muscles after squint, orbital or eyelid surgery, or thyroid eye disease and where scarring of the conjunctiva occurs as may happen after glaucoma surgery or in cicatricial disease, inflammatory disease, for example, pemphigoid, or infective disease, for example, trachoma. A further eye problem associated with the contraction of collagen-comprising tissues for which the methods and medicaments of the present invention may be used is the opacification and contracture of the lens capsule after cataract extraction.

In a preferred embodiment, the protein of the invention can be used for the treatment of burns. Contraction of collagen-comprising tissue, which may also comprise other extracellular matrix components, frequently occurs in the healing of burns. The burns may be chemical, thermal or radiation burns and may be of the eye, the surface of the skin or the skin and the underlying tissues. It may also be the case that there are burns on internal tissues, for example, caused by radiation treatment.

A further aspect of the present invention is the inhibition of the contraction of skin grafts. Skin grafts may be applied for a variety of reasons and may often undergo contraction after application. As with the healing of burnt tissues the contraction may lead to both physical and cosmetic problems. It is a particularly serious problem where many skin grafts are needed as, for example, in a serious burns case.

An associated area in which the medicaments and methods of the present invention are of great use is in the production of artificial skin. To make a true artificial skin it is necessary to have an epidermis made of epithelial cells (keratinocytes) and a dermis made of collagen populated with fibroblasts. It is important to have both types of cells because they signal and stimulate each other using growth factors. A major problem up until now has been that the collagen component of the artificial skin often contracts to less than one tenth of its original area when populated by fibroblasts. MMP inhibitors, for example, collagenase inhibitors may be used to inhibit the contraction to such an extent that the artificial skin can be maintained at a practical size.

Protein of SEQ ID NO:276 (157-15-4-0-B11-CS)

The protein of SEQ ID NO:276, encoded by the cDNA of SEQ ID NO:35, is a variant of a testis-specific isoform of human calpastatin protein (Genseq accession number W19395). The protein of SEQ ID NO:276 contains 2 potential transmembrane segments (position 5 to 25 and position 109 to 129) predicted by the software TopPred II (Claros and von Heijne, *CABIOS applic. Notes*, 10 :685-686 (1994)), and a signal peptide (position 8 : LAVILTLLGLAIL/AI). Like the human calpastatin protein (Genseq accession number W19395), the protein of SEQ ID NO:276 is over-represented in testis.

Calpastatin is a physiological inhibitor of calpains. Calpains, a group of ubiquitous Ca2+ - activated cytosolic proteases, are thought to participate in cytoskeletal remodeling events, cellular adhesion, shape change, and mobility by the site-specific regulatory proteolysis of membrane- and actin-associated cytoskeletal proteins (Beckerle et al., Cell 51:569-577, 1987; Yao et al., Am. J. Physiol. 265(pt. 1):C36-46, 1993; and Shuster et al., J. Cell Biol. 128:837-848, 1995). Calpains have also been implicated in the pathophysiology of cerebral and myocardial ischemia, platelet activation, NF-kB activation, Alzheimer's disease, muscular dystrophy, cataract progression and rheumatoid arthritis. There is considerable interest in inhibitors of calpain, as cellular adhesion, cytoskeletal remodeling events and cell mobility are linked to numerous pathologies (Wang et al., Trends in Pharm. Sci. 15:412-419, 1994; Mehdi, Trends in Biochem. Sci. 16:150-153, 1991). In addition, as the calpain/calpastatin system is involved in membrane fusion events for several cell types, and calpain can be detected in human sperm and testes extracts by Western blotting with specific antisera, tCAST may modulate calpain in the calcium-mediated acrosome reaction that is required for fertilization (Li S et al., Biol Reprod, 63(1):172-8, 2000).

Calpastatin consists of a unique N-terminal domain (domain L) and four repetitive protease-inhibitor domains (domains 1-4) (Lee WJ et al., J Biol Chem, 267(12):8437-42, 1992). The isolated cDNAs from various mammalian species have conspicuous differences in the regions encoding the N-terminal sequences and can be classified into four types. Alternative splicing is most likely the cause for the molecular diversity, and the multiple isoforms are implicated in specific physiological roles (Lee WJ et al., J Biol Chem, 267(12):8437-42, 1992). Type IV (or human tCAST), a shorter isoform, is specifically expressed in testis (Takano J et al., J Biochem Tokyo; 128(1):83-92, 2000). Human tCAST consists of a 40-amino-acid N-terminal T domain plus a part of domain II and all of domains III and IV from the somatic isoform. The protein of SEQ ID NO:276 shows extensive homology to the N-terminal region of the testis basic specific protein (U60665) and the human calpastatin protein (W19395). The homologous region corresponds to domain T and II of the human calpastatin protein (W19395). The T domain targets cytosolic localization and membrane association of tCAST, whereas domain I of somatic calpastatin proteins (sCAST) exhibits a nuclear localization function (Li S et al., Biol Reprod, 63(1):172-8, 2000).

It is believed that the protein of SEQ ID NO:276 is a member of the calpastatin family and, as such, plays a role in cytoskeletal remodeling events, cellular adhesion, shape change, and mobility by the site-specific regulatory proteolysis of membrane- and actin-associated cytoskeletal proteins. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO:276 from positions 1 to 119. Other preferred polypeptides of the invention are any fragments of SEQ ID NO:276 having any of the biological activities described herein.

One embodiment of the present invention relates to methods of using the protein of the invention or part thereof in assays to detect the presence of calpain in a biological sample, such as in bodily fluids, in tissue samples, or in mammalian cell cultures. As calpastatin can bind calpain (Murachi, Biochemistry Int., 18(2)263-294, 1989), the protein of the invention can be used in assays and diagnostic kits to test the presence of calpain using techniques known to those skilled in the art. Preferably, a defined quantity of the protein of the invention or part thereof is added to the sample under conditions allowing the formation of a complex between the protein of the invention or part thereof, and the presence of the complex and/or the free protein of the invention or part thereof is assayed and compared to a control. Calpastatin has been shown to be useful as a marker of intracellular calpain activation, and can be used for monitoring the involvement of calpain in pathological situations (De Tullio et al., FEBS letter, 475(1):17-21, 2000). Calpain has been implicated in cytoskeletal protein degradation involved in the pathophysiology of ischemia and disorders like Alzheimer's disease (Wronski et al., J. Neural transm., 107(2):145-157, 2000), apoptosis in neural cells of rat with spinal cord injury (SCI) (Ray, Brain res., 867(1-2):80-9, 2000), cell fusibility (Kosower et al., Methods Mol Biol., 144:181-94, 2000) and other physiopathologies. Assays detecting any increased calpain level in a cell would thus allow the diagnosis of any of the herein-described diseases or conditions. In addition, a recent study showed that in addition to their proteolytic activities on cytoskeletal proteins and other cellular regulatory proteins, calpain-calpastatin systems can also affect expression levels of genes encoding structural or regulatory proteins (Chen et al., Am. J. Physiol. Cell Physiol, 279:C709-C716, 2000). Thus, the ability to detect calpastatin and calpain levels will likely be useful for the diagnosis of an even larger number of diseases and conditions.

In another embodiment, the polynucleotides or polypeptides of the invention may be used for the detection of gametes, or of specific structures within the gametes, using any technique known to those skilled in the art, including those involving the use of specific antibodies and nucleic acid probes. Various studies have shown that calpastatin is present in the sperm acrosome (Li et al., Bio. of Reprod., 63:172-178, 2000), and more precisely between the plasma membrane and outer acrosomal membrane of cynomolgus macaque spermatozoa (Yudin AI, J Androl, 21(5):721-9, 2000). The ability to visualize spermatozoa generally, or the sperm acrosome in particular, has obvious utility for a number of applications, including for the analysis of infertility in patients, as described below.

Another embodiment of the present invention relates to a method of inhibiting a calpain in a cell. Various studies have shown that it is possible to inhibit calpains dose dependently in cell free protease activity assays: the calpain inhibitor Cerebrolysin can protect microtubule associated protein 2 (MAP2) in a rat model of acute brain ischemia (Wronski et al., J. Neural Transm. Suppl., 59:263-272, 2000), and E-64-D, a cell permeable and selective inhibitor of calpain, can attenuate calpain activity associated with apoptosis in rat SCI (Ray et al., Brain Res., 867(1-2)80-9, 2000). Similarly, it is believed that the protein of SEQ ID NO:276 can be used to inhibit calpain in vitro or in vivo. As calpain has been implicated in a number of pathological processes, diseases, and conditions, such as the pathophysiology of cerebral and myocardial ischemia, platelet activation, NF-kB activation, Alzheimer's disease, muscular dystrophy, cataract progression and rheumatoid arthritis, any of these diseases or conditions can be treated or prevented by increasing or decreasing the activity or expression of the present protein in cells of a mammal affected by the disease or condition. Such an increase can be effected in any of a number of ways, including, but not limited to introducing a polynucleotide encoding the protein of the invention, operably linked to a promoter, into a cell ; and by administering to a cell a compound that increases the activity or expression of the protein of the invention. In addition, the expression or activation of the protein of the invention can be inhibited in any of a large number of ways, including using antisense oligonucleotides, antibodies, dominant negative forms of the protein, and using heterologous compounds that decrease the expression or activation of the protein. Such compounds can be readily identified, e.g. by screening candidate compounds and detecting the level of expression or activity of the protein using any standard assay.

In another preferred embodiment, the protein of the invention can be used to modulate and/or characterize fertility, including for the treatment or diagnosis of infertility, and for contraception. As the calpain/calpastatin system has been implicated in the acrosomal reaction which is a required step in fertilization, it is likely that the over- or under-expression or activation of the present protein disrupts this reaction, thereby inhibiting fertility. Thus, the cause of infertility in many patients can likely be detected by detecting the level of expression of the present protein, where an abnormal level of activity or expression of the protein indicates that a cause of infertility involves the calpain-dependent acrosomal reaction. Such a diagnosis would also point to methods of treating the infertility, e.g. by increasing or decreasing the expression or activation of the protein in spermatozoa. Alternatively, for contraception, the expression or activation of the protein can be artificially disrupted, for example by increasing the protein level using polynucleotides encoding the protein, using the protein itself, or using activators of protein expression or activity, or by decreasing the protein level using inhibitors such as antisense oligonucleotides, antibodies, dominant negative forms of the protein, and using heterologous compounds that inhibit protein expression or activity.

Protein of SEQ ID NO: 295 (internal designation 181-20-3-0-B5-CS)

The protein of SEQ ID NO:295,encoded by the cDNA of SEQ ID NO:54, shows homology to the rat, bovine, and human uromodulin precursor, Tamm-Horsfall urinary glycoprotein, and thuman pancreatic secretory granule membrane major glycoprotein GP2 precursor. SEQ ID NO:295 exhibits homology in the 5' region (over 40% identical and 60% similar) to both GP2 and uromodulin. Like GP2 and uromodulin, the homologous segment contains EGF-like calcium-binding domains, several potential disulfide bonds, and a number of potential N-linked glycosylation sites. Calcium binding EGF-like domains contain a calcium-binding site at the N-terminus, and have been found in proteins which require calcium for their biological activity. Non-limiting examples of proteins which contain calcium-binding EGF-like domains include: (1) Coagulation Factors X, VII, IX; (2) LDL receptors; (3) thrombomodulin; and (4) fibrillin-1. Downing *et al.* [Cell 85:597-605 (1996)] described disease-causing mutations that destabilized a covalently-linked pair of Ca^{2+}-binding EGF-like domains in fibrillin-1 (associated with Marfan Syndrome). These domains form a rigid rod-like arrangement, stabilized by interdomain calcium binding and hydrophobic interaction. Uromodulin (URO) is a 90-100 KDa glycoprotein synthesized by epithelial cells of the ascending loops of Henle and convoluted tubule of the bladder. Except for glycosylation, URO is identical to Tamm-Horsfall protein (THP), the most abundant protein in normal human urine. The relative abundance and specific nephronal location of URO suggests that it may have important physiologic functions in the urinary system.

URO has also been found to be an immunosuppressive glycoprotein, inhibiting antigen-induced human T-cell proliferation. More recent studies have shown that URO can trigger the inflammatory response of neutrophils and stimulate human mononuclear cells to proliferate and release cytokines and gelatinase.

Uromodulin has been shown to play a role in regulating the circulating activity of cytokines since it binds to recombinant interleukin -1 and -2 and tumor necrosis factor (TNF) with high affinity. Although URO does not inhibit the cytotoxic activity of TNFα as monitored by lysis of tumor cell targets, it interacts with recombinant TNFα via carbohydrate chains. This interaction may be critical in promoting clearance and/or reducing *in vivo* toxicity of TNFα and other lymphokines. Endotoxic shock and sepsis are caused by cytokines IL-1 and TNFα. Since URO appears to exhibit inhibitory activity against IL-1 and TNFα, URO may be effective as a therapeutic agent against these conditions. Uromodulin has also been implicated as a possible inhibitor of certain types of bacterial infection in the bladder and urinary tract. URO has the ability to bind to type 1 pilus of *Escherichia coli* and prevent attachment to the surface of epithelium.

SEQ ID NO:295 also has homology to the glycoprotein GP-2. GP-2 is an integral protein of the pancreatic zymogen granule membrane. GP2 is anchored to the lipid bilayer via a glycosyl phosphatidylinositol (GPI) linkage and released by a calcium-activated enzyme into the content of the zymogen granule. Through the process of exocytosis, GP2 is discharged into the pancreatic duct. The protein is also soluble in the zymogens stored in the granule, secreted by the pancreas,

and detected in the pancreatic secretions. GP2 appears to play a role in progression of pancreatitis, an inflammation of the pancreas accompanied by autodigestion of pancreatic tissue by its own enzymes. After cloning and sequencing of GP2, a search of the Genbank database revealed one homologous protein, namely uromodulin. Studies reveal that GP2 and URO not only share structural homology, but functionally are similar in that both can form ductal precipitates under pathological conditions. The aggregation of these precipitates in the pancreas may lead to obstruction of the pancreatic ducts and play a critical role in development of pancreatitis. Similarly, aggregation of URO in the kidney may lead to blockage of the renal tubules and result in renal disease.

The subject invention provides the protein of SEQ ID NO:295 and polynucleotide sequences encoding SEQ ID NO:295. Also included in the invention are biologically active fragments of the protein encoded by SEQ ID NO:295 and polynucleotide sequences encoding these biologically active fragments. "Biologically active fragments" are defined as those peptide or polypeptide fragments of SEQ ID NO:295 which have at least one of the biological functions of the full length protein (e.g., the ability to chelate calcium, bind to *E. coli* pili, or cause immunomodulation of an individual). In one embodiment, the polypeptides of SEQ ID NO:295 are interchanged with the polypeptides encoded by the human cDNA of clone 181-20-3-0-B5-CS.

The invention also provides variants of SEQ ID NO:295. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence of SEQ ID NO:295. Variants according to the subject invention also have at least one functional or structural characteristic of SEQ ID NO:295, such as the biological functions described above or EGF-like calcium-binding domains. The invention also provides biologically active fragments of the variant proteins. Unless otherwise indicated, the methods disclosed herein can be practiced utilizing SEQ ID NO:295, or variants thereof. Likewise, the methods of the subject invention can be practiced using biological fragments of SEQ ID NO:295 , or variants of said biologically active fragments.

Because of the redundancy of the genetic code, a variety of different DNA sequences can encode SEQ ID NO:295. It is well within the skill of a person trained in the art to create these alternative DNA sequences which encode proteins having the same, or essentially the same, amino acid sequence. These variant DNA sequences are, thus, within the scope of the subject invention. As used herein, reference to "essentially the same sequence" refers to sequences that have amino acid substitutions, deletions, additions, or insertions that do not materially affect biological activity.

"Recombinant nucleotide variants" are alternate polynucleotides which encode a particular protein. They can be synthesized, for example, by making use of the "redundancy" in the genetic code. Various codon substitutions, such as the silent changes which produce specific restriction sites or codon usage-specific mutations, can be introduced to optimize cloning into a plasmid or viral vector or expression in a particular prokaryotic or eukaryotic host system, respectively.

SEQ ID NO:295, and variants thereof, can be used to produce antibodies according to methods well known in the art. The antibodies can be monoclonal or polyclonal. Antibodies can also be synthesized against fragments SEQ ID NO:295 as well as variants of SEQ ID NO:295 according to known methods. The subject invention also provides antibodies which specifically bind to biologically active fragments of SEQ ID NO:295 or biologically active fragments of variants of SEQ ID NO:295.

The subject invention also provides for immunoassays which are used to screen for, monitor, or diagnose conditions or disorders associated with liver dysfunction and/or damage. These conditions or disorders include, and are not limited to, hepatitis, cirrhosis, fibrosis, pericholangitis, portal triaditus, chronic periportal inflammation, systemic lupus erythematosus, Hodgkin's disease, Granulomas, and cell dysplasia can also be diagnosed. For a number of disorders listed above, expression of these genes at significantly higher or lower levels can be routinely detected in certain liver tissues or cell types (e.g., cancerous) or bodily fluids (e.g., serum, plasma, and blood) taken from an individual having such a disorder, relative to the standard gene expression levels, e.g., the expression level in healthy tissue or bodily fluid from one or more individuals not having the disorder. These types of assays allow for a non-invasive method of screening for, diagnosing, or monitoring liver cancer in human subjects. Similarly, antibodies and small molecules directed to the polypeptides can be used as immunological probes for differential identification of the diseased tissue(s) or cells.

Additionally, nucleic acid and amino acid sequences of SEQ ID NOs:54 and 295 can be used to provide polypeptides and biologically active fragments thereof for the repair of cellular injury following liver damage and/or liver transplant.

Furthermore, polypeptides, or biologically active fragments thereof, can be used for the modulation of bacterial binding to epithelial cells or as a modulator of bacterial infection. In this aspect of the subject invention, bacterial cells are contacted with an amount of a composition comprising the polypeptide, or biologically active fragments thereof, sufficient to interfere with the binding of bacteria to epithelial cells. In one embodiment, the bacteria are coliform bacterial cells. In another embodiment, the bacterial cells are *E. coli*. Compositions comprising SEQ ID NO:295, or biologically active fragments thereof, can be administered in any fashion required to provide a therapeutic effect (e.g., orally, intravenously, intrathecally, intraarterially, etc.).

The subject invention also provides materials and methods for the treatment of endotoxic shock and/or sepsis. In this embodiment, a subject can be treated with therapeutically effective amounts of a composition comprising SEQ ID NO:295, or biologically active fragments thereof.

The subject invention also provides materials and methods for the *in vivo* or *in vitro* chelation of calcium ions (Ca^{2+}). In this aspect of the invention, SEQ ID NO: , or biologically active fragments thereof, can be used to bind free Ca^{2+} by addition of the polypeptide, or biologically active fragments thereof, to solutions, environmental samples, or biological samples.

Alternatively, a composition containing the SEQ ID NO:295, or biologically active fragments thereof, can be added to the solutions, environmental samples, or biological samples in amounts sufficient to bind and remove free Ca^{2+} from solution.

In another aspect of the subject invention, SEQ ID NO:295, or biologically active fragments thereof, can be used to modulate the immune system of a mammal. In this method, immunomodulatory amounts of SEQ ID NO:295, or biologically active fragments thereof, can be administered to a mammal in a pharmaceutically acceptable carrier. Methods of assessing the stimulated state of the immune system of the mammal can be practiced according to methods well known in the art.

Protein of SEQ ID NOs:244, 251 (internal designation numbers 105-016-3-0-G10-CS and 105-074-3-0-H10-CS)

The 274 amino acid protein of SEQ ID NO:244, encoded by the cDNA of SEQ ID NO:3, found in prostate and strongly expressed in the salivary gland, presents strong sequence similarities with the yeast putative mitochondrial carrier protein PET8 (SWISSPROT accession number P38921) and with similar proteins conserved among eukaryotes (D. *melanogaster* and C. *elegans*: respective SPTREMBLNEW SPTREMBL SWISSPROT accession numbers Q9VBN7 and Q18934, and S. *pombe*: SWISSPROT accession number: Q10442). All members of the mitochondrial carrier/transport protein superfamily exhibit sequence motifs highly similar to P-X-D/E-X-X-K/R-X-R that are also found in 3 positions in the protein of the invention (positions 26 to 33, 108 to 115 and 199 to 206) (Belenkiy *et al*, Biochim. Biophys. Acta, 1467:207-218 (2000)). These mitochondrial carrier protein signatures are associated with membrane-spanning segments (Belenkiy *et al*, ibid; Kuan et Saier, Crit. Rev. Biochem. Mol. Biol., 28:209-233 (1993)). In fact, 4 candidate membrane-spanning segments are identified in the protein of the invention, from amino acid positions 4 to 24, 51 to 71, 180 to 200 and 240 to 260. Other hydrophobic regions are found in positions 86 to 107 and 139 to 162. In addition, the protein of SEQ ID NO:244 presents a putative signal peptide in its very amino-terminal part (position 5 to 19).

The protein of SEQ ID NO:251, encoded by the cDNA of SEQ ID NO:10, is a 72 amino acid truncated form of the protein of SEQ ID NO:244. This shorter product results from the absence, in the cDNA of SEQ ID NO:10, of the 110bp exon (position 275 to 384) found in the cDNA of SEQ ID NO:3. Nevertheless, the 72 amino acid encoded protein possesses the putative signal peptide (position 5 to 19), the first mitochondrial carrier protein signature (position 26 to 33), and two candidate membrane-spanning segments (positions 4 to 24 and 51 to 71).

Energy transduction in mitochondria requires the transport of many specific metabolites across the inner membrane of this eukaryotic organelle. Different types of substrate carrier proteins involved in energy transfer are found in the inner membrane. These proteins all seem to be evolutionary related, and constitute the mitochondrial carrier/transport proteins (MCP/MTP)

superfamily. Structurally, MCP/MTP proteins are typically homodimeric integral transmembrane polypeptides (subunit molecular weight ~30kD) that traverse the inner mitochondrial membrane six times with both the N- and C-termini localized to the cytosolic side of the membrane. Each 30kD subunit is composed of three tandem repeats of a domain of approximately one hundred residues (~10kD). This 10kD domain contains two transmembrane regions and a sequence motif highly similar to P-X_1-D/E-X_2-X_3-K/R-X_4-R, where X_3 is a hydrophobic residue (Kuan et Saier, Crit. Rev. Biochem. Mol. Biol., 28:209-233 (1993)). Five protein families of known function have been identified among the mitochondrial carrier protein superfamily:

(1) The ADP, ATP carrier protein (ACC), ADP/ATP translocases, which under the conditions of oxidative phosphorylation catalyze the one to one exchange of cytosolic ADP against matrix ATP across the inner mitochondrial membrane (Fiore *et al*, Biochimie, 80:137-150 (1998)). The ADP/ATP transport system can be blocked very specifically by two families of inhibitors: atractyloside (ATR) and carboxyatractyloside (CATR) on one hand, and bongkrekic acid (BA) and isobongkrekic acid (isoBA) on the other hand. It is well established that these inhibitors recognise two different conformations of the carrier protein, the CATR- and BA-conformations, which exhibit different chemical, immunochemical and enzymatic reactivities. Bakker and collaborators have reported that myopathies might result from a defect in ADP/ATP transport (Bakker *et al*, Pediatr. Res. 33:412-417 (1993)). Namely, the authors describe a 4-fold decrease in the concentration of the ADP/ATP carrier protein in a patient with a mitochondrial myopathy.

(2) The 2-oxoglutarate/malate carrier protein (OGCP), which exports 2-oxoglutarate into the cytosol and imports malate, or other dicarboxylic acids, into the mitochondrial matrix. This protein plays an important role in several metabolic processes, such as the malate/aspartate and the oxoglutarate/isocitrate shuttles (Palmieri *et al*, J. Bioenerg. Biomembr. 25:493-501 (1993)).

(3) The phosphate group carrier protein, which transports inorganic phosphate groups from the cytosol into the mitochondrial matrix (Palmieri *et al*, ibid).

(4) The mammalian brown fat uncoupling proteins, such as UCP-1 (thermogenin), are transmembrane proton-translocating proteins present in the mitochondria of brown adipose tissue, a specialized tissue which functions in heat generation and energy balance ((Jezek and Garlid, Int. J. Biochem. Cell. Biol. 30:1163-1168 (1998); Klingenberg, J. Bioenerg. Biomembr. 31:419-430 (1999); Nicolls and Locke, Physiol. Rev. 64:2-40 (1994); Rothwell and Stock, Nature 281:31-35 (1979)). Mitochondrial oxidation of substrates is accompanied by proton transport out of the mitochondrial matrix, creating a transmembrane proton gradient. Typically, re-entry of protons into the matrix via ATP synthase is coupled to ATP synthesis. However, UCP-1 functions as a transmembrane proton transporter, permitting re-entry of protons into the mitochondrial matrix unaccompanied by ATP synthesis. Environmental exposure to cold evokes neural and hormonal stimulation of brown adipose tissue, which increases UCP mediated proton transport, brown fat metabolic activity, and heat production.

Studies with transgenic models indicate that brown fat and UCP-1 play an important role in energy expenditure in rodents. Transgenic mice in which brown adipocyte tissue was ablated by a toxin coupled to the UCP-promoter developed obesity and diabetes (Lowell *et al.*, Nature 366:740-742 (1993)). Obesity in these transgenic animals developed in the absence of hyperphagia, suggesting that the uncoupled mitochondrial respiration of brown fat is an important component of energy expenditure. In a separate transgenic mouse model, ectopic expression of UCP-1 in white adipose tissue of genetically-obese mice led to a significant reduction in body weight and fat stores (Kopecky *et al.*, J. Clin. Invest. 96:2914-2923 (1995)). These studies indicate that activity of UCP-1 is accompanied by energy expenditure and weight loss in rodents. Two other UCP proteins have recently been cloned. The first uncoupling protein-like protein (UCPL) or UCP-2 (59% homologous), is widely expressed (heart, kidney, lung, placenta and white fat) and enriched in tissues of the lymphoid lineage (Fleury *et al.*, Nature Genetics 15:269-272 (1997)). The second, UCP-3 3 (57% homologous), is predominantly localized to skeletal muscle and brown fat (Boss et *al.*, FEBS Lett. 408:39-42 (1997)). UCP-3 has been found to be regulated by cold and thyroid hormone (Larkins *et al.*, Biochem. Biophys. Res. Comm. 240:222-227 (1997)).

Thermogenic protein activity, such as that found with UCP-1, may be useful in reducing, or preventing the development of, excess adipose tissue, such as that found in obesity. Obesity is becoming increasingly prevalent in developed societies. Attempts to reduce food intake, or to decrease hypernutrition, are usually fruitless in the medium term because the weight loss induced by dieting results in both increased appetite and decreased energy expenditure (Leibel *et al.*, New Engl. J. Med. 322:621-628 (1995)). The intensity of physical exercise required to expend enough energy to materially lose adipose mass is too great for many obese people to undertake on a sufficiently frequent basis. Thus, obesity is currently a poorly treatable, chronic, essentially intractable metabolic disorder. In addition, obesity carries a serious risk of co-morbities including, Type 2 diabetes, increased cardiac risk, hypertension, atherosclerosis, degenerative arthritis, and increased incidence of complications of surgery involving general anesthesia.

(5) The tricarboxylate transport protein (or citrate transport protein), which is involved in citrate-H+/malate exchange. This protein is important for the bioenergetics of hepatic cells as it provides a carbon source for fatty acid and sterol biosyntheses, and NAD for the glycolytic pathway (Kaplan *et al*, J. Biol. Chem. 268:13682-13690 (1993)).

It is believed that the protein of SEQ ID NO:244 or part thereof is a member of the mitochondrial carrier/transport protein superfamily and, as such, plays a role in mitochondrial processes such as ADP/ATP, malate/aspartate, 2-oxoglutarate/isocitrate, citrate-H+/malate exchanges across the inner membrane, phosphate groups transport and physiological roles such as regulation of body weight and energy balance, muscle nonshivering thermogenesis, fever, and defense against the generation of reactive oxygen species. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO:244 from amino acid positions 26 to

33, 108 to 115 and 199 to 206 on one hand, and from positions 4 to 24, 51 to 71, 86 to 107, 139 to 162, 180 to 200 and 240 to 260 on the other hand. Other preferred polypeptides of the invention are fragments of SEQ ID NO:244 having any of the biological activities described herein. It is believed that the protein of SEQ ID NO:251 is a 72 amino acid truncated form of the 274 amino acid protein of SEQ ID NO:244, and corresponds to one subunit of the tripartite structure of mitochondrial carrier/transport proteins. Preferred polypeptides are polypeptides comprising the amino acids of SEQ ID NO:251 from positions 4 to 24, 26 to 33 and 51 to 71.

The activity of the protein of the invention can be assessed using cultured cells. For example, nucleic acids encoding the protein of SEQ ID NO:244 can be cloned into a eukaryotic vector and transfected into a population of cells. Transfected mammalian cells are then tested for their carrier activity e.g., the import of ADP, dicarboxylic acids, inorganic phosphate groups, or H^+ into the mitochondrial matrix, and the export of ATP, 2-oxoglutarate, tricarboxylate-H^+ export into the cytosol. These transfected cell lines may allow the development of *in vitro* assays for the identification of modulators of the carrier activity, such as atractyloside (ATR), carboxyatractyloside (CATR), bongkrekic acid (BA) and isobongkrekic acid (isoBA), which were described above in connection with the ADP/ATP mitochondrial carrier. Such modulators are useful for the treatment of any diseases or conditions associated with the protein of the invention.

Another embodiment of the invention relates to compositions and methods using the protein of the invention or part thereof to label mitochondria, or more specifically the inner mitochondrial membrane, in order to visualize any change in number, topology or morphology of this organelle, for example in association with a mitochondria-related human disorder, such as neuroleptic malignant syndrome (NMS) (Kubo et al., Forensic Sci. Int. 115:155-158 (2001)), the Rett syndrome (Armstrong, Brain Dev. 14 Suppl:S89-98 (1992)), Alpers disease (Chow and Thorburn, Hum. Reprod. 15 Suppl 2:68-78 (2000)) or mitochondrial encephalomyopathies (Handran et al., Neurobiol. Dis. 3:287-298 (1997)). For example, the protein may be rendered easily detectable by inserting the cDNA encoding the protein of the invention into a eukaryotic expression vector in frame with a sequence encoding a tag sequence. Eukaryotic cells expressing the tagged protein of the invention may also be used for the *in vitro* screening of drugs or genes capable of treating any mitochondria-related disease or conditions. The protein of the invention can also be used to specifically label cells of the salivary gland or of the prostate, e.g. for histological analyses or for the identification of the origin of tumor cells.

The protein of the invention can also be used as a carrier/transporter to translocate radiolabeled or chemically labeled metabolites (ADP, dicarboxylic acids, inorganic phosphate groups) from the cytosol to the matrix of the mitochondria in order to specifically label this organelle, e.g. to follow its modifications. For example, radiolabeled

or chemically labeled precursors can be added to an *in vitro* culture of mammalians cells stably transfected and expressing the protein of the invention. The labeling of the organelles can then be stopped at different times after the beginning of the experiment by adding specific inhibitors of carrier/transporter proteins, such as atractyloside (ATR), carboxyatractyloside (CATR), bongkrekic acid (BA), or isobongkrekic acid (isoBA). Cells with labeled mitochondria can be used for the *in vitro* screening of drugs or genes capable of causing mitochondrial modifications.

Still another embodiment of the invention or part thereof relates to methods of delivering heterologous compounds, either polypeptides or polynucleotides, to the inner membrane of mitochondria by recombinantly or chemically fusing a fragment of the protein of the invention to a heterologous polypeptide or polynucleotide. Preferred fragments are the putative peptide signal, the four membrane-spanning segments and/or any other fragments of the protein of the invention that may contain targeting signals for mitochondria including but not limited to matrix targeting signals as defined in Herrman and Neupert, Curr. Opinion Microbiol. 3:210-4 (2000); Bhagwat et al. J. Biol. Chem. 274:24014-22 (1999), Murphy Trends Biotechnol. 15:326-30 (1997); Glaser et al. Plant Mol Biol 38:311-38 (1998); Ciminale et al. Oncogene 18:4505-14 (1999). Such heterologous compounds may be used to modulate mitochondrial activities, such as to induce and/or prevent mitochondrial-induced apoptosis or necrosis. For example, these heterologous compounds may be used in the treatment and/or the prevention of disorders in which apoptosis is deleterious, including, but not limited to, immune deficiency syndromes (including AIDS), type I diabetes, pathogenic infections, cardiovascular and neurological injury, alopecia, aging, degenerative diseases such as Alzheimer's Disease, Parkinson's Disease, Huntington's disease, dystonia, Leber's hereditary optic neuropathy, schizophrenia, and myodegenerative disorders such as "mitochondrial encephalopathy, lactic acidosis, and stroke" (MELAS), and "myoclonic epilepsy ragged red fiber syndrome" (MERRF). In addition, heterologous polynucleotides may be used to deliver nucleic acids for mitochondrial gene therapy, i.e. to replace a defective mitochondrial gene and/or to inhibit the deleterious expression of a mitochondrial gene.

The invention further relates to methods and compositions used to modify the protein of the invention. Post-translational modifications encompassed by the invention include, N-linked or O-linked carbohydrate chains, processing of N-terminal or C-terminal ends, attachment of chemical moieties to the amino acid backbone, chemical modifications of N-linked or O-linked carbohydrate chains, and addition or deletion of an N-terminal methionine residue as a result of prokaryotic host cell expression. These post-translational modifications of the protein of the invention may be very

useful in the search for its putative protein partners, using approaches such as screening of an expression cDNA library with a radiolabeled recombinant protein, as post-translational modifications are of first importance in protein-protein interactions. Identification of proteinic partners of mitochondrial carrier proteins would allow the study of their regulation *ex vivo* and *in vivo* in normal versus pathologic cases (for an example concerning the UCP1 mitochondrial carrier protein and its 14.3.3 physical partner, see: Pierrat et al., Eur. J. Biochem. 267:2680-2687 (2000)).

Another embodiment of the invention relates to composition and methods using polynucleotide sequences encoding the protein of the invention or part thereof to establish transgenic model animals (D. *melanogaster*, M. *musculus*), by any method familiar to those skilled in the art. By modulating *in vivo* the expression of the transgene with drugs or modifier genes (activator or suppressor genes), animal models can be developed that mimic human mitochondria-associated disorders such as myopathies or obesity. These animal models would thus allow the identification of potential therapeutic agents for treatment of the disorders. In addition, recombinant cell lines derived from these transgenic animals may be used for similar approaches *ex vivo*.

In one embodiment, the protein of SEQ ID NO:251, corresponding to the 72 amino acid truncated form of SEQ ID NO:244, may be used as a dominant negative variant to inhibit the function of the full-length form of the protein of SEQ ID NO:244 in vitro or in vivo. Inactivation of mitochondrial carriers in this way may allow the development of animal models for human disorders. Recently, for example, Lowell and collaborators have shown in the mouse that a targeted destruction of UCP1 by the diphteria toxin A chain is able to produce obese animals (Kozak and Koza, ibid., Lowell et al., ibid.).

Protein of SEQ ID NO: 285 (internal designation 174-39-2-0-A3-CS)

The protein of SEQ ID NO:285, encoded by the cDNA of SEQ ID NO: 44 (clone 174-39-2-0-A3-CS), is overexpressed in cancerous prostate, fetal brain, muscle and placenta. The protein is homologous to the NADH-cytochrome b5 reductase isoform and to the human electron transport protein.

NADH-cytochrome b5 reductase proteins belong to a flavoenzyme family sharing common structural features and whose members (ferrodoxin-NADP+ reductase, NADPH-cytochrome P450 reductase, NADPH-sulfite reductase, NADH-cytochrome b5 reductase and NADH-nitrate reductase) are involved in photosynthesis, in the assimilation of nitrogen and sulfur, in fatty-acid oxidation, in the reduction of methemoglobin and in the metabolism of many pesticides, drugs and carcinogens (Karplus et al., Science, 251:60-6 (1991)). In addition, cytochrome b5 reductase is thought to play a role in the prevention of apoptosis following oxidative stress (see review by Villalba et al., Mol Aspects Med 18 Suppll:S7-13 (1997)).

It is believed that the protein of SEQ ID NO: 285 may be an oxidoreductase. Thus it may play a role in electron transport and general aerobic metabolism and may be associated with mitochondrial

membranes. In addition, the protein of the invention may be able to use FAD and/or molybdopterin as cofactors. It may be involved in photosynthesis, in the assimilation of nitrogen and sulfur, in fatty-acid oxidation, in the reduction of methemoglobin and in the metabolism of many pesticides, drugs and carcinogens. Preferred polypeptides of the invention are fragments of SEQ ID NO: 285 having any of the biological activity described herein. The oxidoreductase activity of the protein of the invention may be assayed using any technique known to those skilled in the art. The ability to bind a cofactor may also be assayed using any techniques well known to those skilled in the art including, for example, the assay for binding NAD described in US patent 5,986,172.

In another embodiment, the protein of the invention or part thereof is used to prevent cells from undergoing apoptosis. In a preferred embodiment, the apoptosis active polypeptide is added to an in vitro culture of mammalian cells in an amount effective to reduce apoptosis. Furthermore, the protein of the invention or part thereof may be useful in the diagnosis, the treatment and/or the prevention of disorders in which apoptosis is deleterious, including but not limited to immune deficiency syndromes (including AIDS), type I diabetes, pathogenic infections, cardiovascular and neurological injury, alopecia, aging, degenerative diseases such as Alzheimer's Disease, Parkinson's Disease, Huntington's disease, dystonia, Leber's hereditary optic neuropathy, schizophrenia, and myodegenerative disorders such as "mitochondrial encephalopathy, lactic acidosis, and stroke" (MELAS), and "myoclonic epilepsy ragged red fiber syndrome" (MERRF).

The invention further relates to methods and compositions using the protein of the invention or part thereof to diagnose, prevent and/or treat several disorders in which energy metabolism is impaired, or needs to be impaired, including but not limited to mitochondriocytopathies, necrosis, aging, neurodegenerative diseases, myopathies, methemoglobinemia, hyperlipidemia, obesity, cardiovascular disorders and cancer. For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. For prevention and/or treatment purposes, the protein of the invention may be used to enhance electron transport and increase energy delivery using any of the gene therapy methods described herein.

Protein of SEQ ID NO:368 (internal designation 187-45-0-0-118-CS)

The protein of SEQ ID NO: 368 encoded by the cDNA of SEQ ID NO: 127 is a 78 amino acids long polypeptide. The sequence of the protein of SEQ ID NO: 368 is identical to the sequence of the human Dad1 protein, the defender against apoptotic cell death 1 protein, a subunit of the mammalian oligosaccharyltransferase (OST), except that the last 43 residues of Dad1 are replaced by a series of 8 different amino acids in the protein of the invention. In addition, the protein of SEQ ID NO: 368 displays the pfam signature for DAD family proteins from positions 1 to 78 as well as two putative transmembrane domains from positions 31 to 51 and 54 to 74. The Dad1 protein is a 113 amino acids long protein which mRNA is composed of 3 exons [see Genbank accession

number D15057 and Nakashima, T. *et al* (1993) *Molecular and Cellular Biology* 13:6367-6374]. The cDNA of SEQ ID NO: 127 is composed of the first and third exon of the Dad1 cDNA wheras the second exon of the Dad1 cDNA is missing. Taken together, these datas indicate that the protein of SEQ ID NO: 127 is a new isoform of the Dad1protein resulting from an alternative splicing event.

Asparagine-linked glycosylation is a highly conserved protein modification reaction that occurs in all eukaryotes. The initial stage in the biosynthesis of N-linked glycoproteins, catalysed by the enzyme oligosaccharyltransferase (OST), involves the transfer of a preassembled high-mannose oligosaccharide from a dolichol-linked oligosaccharide donnor onto asparagine acceptor sites in nascent proteins in the lumen of the rough endoplasmic reticulum [for review, see Silberstein, S. *et al* (1996) *FASEB J* 10: 849-858].

Protein glycosylation is essential for the structure and function of many proteins and is involved in the control of many diverse biological processes (Paulson, *Trends in Biol. Sci.*, 1989, 14, 272; Sadler, *In Biology of Carbohydrates*, 2nd Ed., Ginsburg & Robbins, Ed., John Wiley & Sons: New York, 1984, Vol. 2, pg. 87). For example, protein glycosylation has been found to be crucial for the development, growth and proper function of complex organisms, while the aberrant glycosylation of proteins has been associated with diseased and transformed cells.

The mammalian oligosaccharyltransferase is composed of the four ER membrane proteins, ribophorin I and II (RI and RII), OST48, and DAD1, which form an oligomeric complex. RI and OST48, and probably also RII, are type I transmembrane proteins. The luminal domain of OST48 interacts with those of RI and RII and the cytoplasmic domain of OST48 has affinity for the cytoplasmically exposed N-terminal tail of DAD1 [Kelleher, D. *et al.* (1997) *Proc Natl. Acad. Sci. USA* 94: 4994-4999; Fu, J. *et al.* (1997) *J. Biol. Chem* 272: 29687-29692].

Dad1 is a small hydrophobic protein, thought to be an integral membrane protein, with a cytoplasmically located N terminus and up to three transmembrane domains. As is true for the other subunits of OST, the precise role of Dad1 in N-glycosylation is not known. However, it has been shown that Dad1 is critical for the function and the structural integrity of the OST complex [Sanjay, A. *et al.* (1998) *J. Biol. Chem* 273: 26094-26099]. Also, it is worth noting that the Dad1 protein was first identified in 1993 as a mammalian cell death suppressor since loss of its function induces apoptosis in hamster BHK21 cells [Nakashima, T. *et al* (1993) *Molecular and Cellular Biology* 13:6367-6374]. Since then, several reports have confirmed the anti-apoptotic role for Dad1 [Hong, NA. *et al*, (2000) *Dev Biol* 220:76-84; Brewster, JL. *et al*, (2000) *Genesis* 26: 271-8); Yoshimi, M. *et al*, (2000) *Biochem Biophys Res Commun* 276: 965-9].

Dad1 is a highly conserved protein whose sequence has been determined for diverse organisms including several vertebrates, a nematode, and several plants. A comparaison of these sequences reveals that the amino-terminal region preceding the first membrane-spanning segment is the least conserved region of the protein both with repect to lengh and amino acid sequence identity.

The most highly conserved sequences of Dad1 include the second and third membrane spanning segments, making them probably the most crucial regions for Dad1 function [Kelleher, D. *et al.* (1997) *Proc Natl. Acad. Sci. USA* 94: 4994-4999]. The importance of the C-terminus region for mediating Dad1 functions has been recently confirmed [Makishima, T. *et al.* (2000) *J. Biochem (Tockyo)* 128:399-405]

Therefore, Dad1 is thought to act as a positive regulator of the oligosaccharyltransferase complex, and as a negative regulator of apoptosis. In addition, the C-terminus of the Dad1 protein seems to be important for mediating these functions. As mentionned above, the protein of the invention is a new isoform of the Dad1 protein resulting from an alternative splicing event. As a result of this alternative splicing event, the C-terminus of the protein of the invention is shortened and does not display the third transmembrane domain of Dad1. Since the C-terminus of Dad1 has been shown to be important for mediating the protein function, it is believed that the protein of the invention has rather an antagonistic action to the one of Dad1. It is worth noting that this type of situation in which the same gene give rise by alternative splicing to different protein products with opposing functions is a commun theme among apoptosis genes [For a review, see Reed, JC. (1999) *Nat. Biotechnol* 17: 1064-65].

Thus, it is believed that the protein of the invention of SEQ ID NO: 368 plays a role in the control of N-glycosylation of celullar proteins. Preferably, the protein of the invention is thought to act as a positive regulator of apoptosis and a negative regulator of the OST complex. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 368 from positions 1 to 78, and 71 to 78. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 368 having any of the biological activity described herein. The activity of the protein of the invention or part thereof on protein N-glycosylation may be assayed using any of the assays known to those skilled in the art. For example, one could use DNA-mediated gene transfer techniques in order to introduce the cDNA sequence of SEQ ID NO: 127 or part thereof into cell lines so that the protein of SEQ ID NO: 368 or part thereof is over expressed in these cell lines. The resulting effect of this over expression on the N-glycosylation of proteins can then be studied using immunoblotting or Western blotting of glycoproteins [Makishima *et al.* (1997) *Genes Cells* 2: 129-141; Silberstein *et al.* (1995) *J. Cell. Biol:* 131: 371-383; Hong *et al.* (2000) *Developmental Biology* 220: . The activity of the protein of the invention or part thereof on cellular apoptosis may be assayed using any of the assays known to those skilled in the art including those described by Nakashima *et al.* (1993) *supra.*

One object of the present invention are compositions and methods of targeting heterologous polypeptides to the endoplasmic reticulum by recombinantly or chemically fusing a fragment of the proteins of the invention to an heterologous polypeptide. Preferred fragments are any fragments of the proteins of the invention, or part thereof, that may contain targeting signals for the endoplasmic reticulum such as those described in Pidoux AL, Armstrong EMBO J 1992 Apr;11(4):1583-91; Munro

S, Pelham HR Cell 1987 Mar 13;48(5):899-907; Pelham HR Trends Biochem Sci 1990 Dec;15(12):483-6.

In another embodiment, the invention relates to compositions and methods using the protein of the invention or part thereof to stimulate cells'entry into apoptosis. In a preferred embodiment, the pro-apoptosis protein of the invention or part thereof is added to an in vitro culture of mammalian or plant cells in an amount effective to stimulate apoptosis. In another preferred embodiment, the cDNA sequence of SEQ ID NO: 127 or part thereof may be used to create transgenic animals or plant cells in which the disclosed protein of the invention or part thereof can be expressed at higher levels than normal whenever and wherever it is desired. Ways to create transgenic cells in which the expression of the transgene can be turn on or off whenever it is desired are well known in the art. Increasing the expression level of the protein of the invention in cells to stimulate programmed cell death may be useful for applications in which a given species of cells become undesirable upon a given event, i.e., infection, transformation, end of a production process, etc...

Furthermore, the invention relates to methods and compositions using the protein of the invention or part thereof to diagnose, prevent and/or treat disorders characterized by abnormal cell proliferation and/or programmed cell death, including but not limited to cancer, immune deficiency syndromes (including AIDS), type I diabetes, pathogenic infections, cardiovascular and neurological injury, alopecia, aging, degenerative diseases such as Alzheimer's Disease, Parkinson's Disease, Huntington's disease, dystonia, Leber's hereditary optic neuropathy, schizophrenia, and myodegenerative disorders such as "mitochondrial encephalopathy, lactic acidosis, and stroke" (MELAS), and "myoclonic epilepsy ragged red fiber syndrome" (MERRF). For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. For prevention and/or treatment purposes of disorders in which cell proliferation needs to be reduced and/or apoptosis increased, the expression of protein of the invention may be enhanced using any of the gene therapy methods described herein or known to those skilled in the art. For prevention and/or treatment purposes of disorders in which cell proliferation needs to be enhanced and/or apoptosis reduced, inhibition of endogenous expression of the protein of the invention may be achieved using any methods or known to those skilled in the art including the triple helix and antisense strategies described herein.

Moreover, antibodies to the protein of the invention or part thereof may be used for detection of the endoplasmic reticulum for histological purposes using any techniques known to those skilled in the art.

Protein of SEQ ID No: 284 (internal designation 174-38-3-0-C9-CS)

The protein of SEQ ID No: 284 encoded by the cDNA of SEQ ID No: 43 is overexpressed in salivary gland. The 406-amino-acid-long protein of invention, which is similar in size to fucosyltransferases, displays a Pfam motif of the fusosyltransferase family from residues 70 to 406. Furthermore, the present protein of invention is homologous to a putative fucosyltransferase of Drosophila melanogaster (STR accession number: Q9VLC1 and Q9VLC1). The protein of SEQ ID 284 also shares homology with the alpha1,3 fucosyltransferase (E.C. 2.4.1.152), found in Brachydanio renio (EMBL accession number : AB023627), Schistosoma mansoni (GENPEPT accession number : AF183577-1), cattle (SPTREMBL accession number Q9TQQ3), and human species (GENPEPT accession number : AJ132772_2). Like fucosyltransferases, the protein of the invention displays the features of type II transmembrane proteins with a short N -terminal cytoplasmic tail, a 9-29 amino acid signal-anchor transmembrane domain, and a large C-terminal domain. Furthermore, the present protein of invention displays an almost perfect consensus motif of the alpha-1,3 fucosyltransferases from residues 315 to 345 (Breton et al. Glycobiology 1998; 1: 87-94).

Fucosyltransferases are a family of enzymes that catalyze the transfer of fucose from GDP-fucose, to galactose in an alpha1,2 linkage, and to N-acetylglucosamine in alpha1,3-, alpha1,4- and alpha1,6- linkages. Since all fucosyltransferases use the same nucleotide sugar, their specificity will probably reside in the recognition of the acceptor and in the type of linkage formed. In human species, fucosyltransferases, which are type II membrane proteins found in Golgi, can be split into three distinct families (Breton et al. Glycobiology 1998; 1: 87-94): (1) the alpha-1,2-fucosyltransferases, hFUT1 and hFUT2, which yield nearly identical products as only single carbohydrate linkage differentiates type I from type II glycans. hFUT1 determines the expression of O-type antigen (H antigen) of the ABO blood group system on erythrocytes, whereas hFUT2 (Se) determines it in saliva, *i.e.* secretor status; (2) The alpha-1,3-fucosyltransferases that constitute a distinct homogenous family of proteins, although some regions display similarities with the alpha-1,2 and alpha-1,6-fucosyltranferases (Breton et al. Glycobiology 1998;1:87-94). Five alpha -1,3-fucosyltransferases have been characterized to date in the human species, *i.e.* hFUT3 (Lewis enzyme), hFUT4 (myeloid-type), hFUT5, hFUT6 (plasma-type), and hFUT7. These are involved in the lasts steps of the biosynthesis the carbohydrate antigen sialyl Lewis of ABH (de Vries et al. J Biol Chem 1995;270:8712-22 ; Kimura et al. Biochem Biophys Res Commun 1997 8;237:131-7) ; (3) The alpha-1,6-fucosyltransferase, hFUT8, which is implicated in the synthesis of N-glycans (Miyoshi et al. Biochim Biophys Acta 1999;1473:9-20).

The fucosylated cell surface glycoconjugates play important roles in physiological and pathological processes, such as fertilization, embryogenesis, lymphocyte trafficking, immune response, and cancer metastasis (Staudacher et al. Trends Glycosci Glycotechnology 1996; 8:391-408). More specifically, the fucosylated cell surface glycoconjugates, which are present on the

apical surface of various epithelium, contribute to resistance of various microorganisms agents including bacteria as Helicobacter pylori (Umesaki et al. Science 1997;276:964-5), and E. coli (Vogeli et al. Schweiz Arch Tierheilkd 1997;139:479-84), and virus such as HIV (Ali et al. Infect Dis 2000;181:737-9). On the other hand, abnormal upregulation of fucosyltransferases is a common finding in various types of tumors, which cause an increased production of fucosylated glycoconjugates. Such fucosylated glycoconjugates can also serve as tumor markers and include (1) the Ca19-9 cancer antigen, which circulating sialyl-Lewis a structure produced by hFUT3 and used for diagnosis of pancreatic and gastric cancer (Koprowski et al. Somatic Cell Genet 1979;5:957-71), and (2) alpha-foetoprotein whose alpha 1,6-fucosylatation is reduced in hepatoma (Miyoshi et al. Biochim Biophys Acta. 1999;1473:9-20).On the other hand, aberrant production of fucosylated glycoconjugates can provide selective growth advantage by facilitating the extravasation of tumor cells, since they participate to endothelial adhesion through interaction with E- and P- selectins of endothelial cells (Butcher and Picker, Science 1996;272:60-6). Consequently, modulation of fucosyltransferase activity can modify tumorogenicity in various model of tumors including hepatoma (Miyoshi et al. Biochim Biophys Acta. 1999;1473:9-20), and colorectal adenocarcinoma (Weston et al. Cancer Res 1999;59:2127-35).

Thus, it is believed that the protein of the invention of SEQ ID NO: 284 is a glycosyltransferase, preferably an hexosyltransferase, more preferably a fusosyltransferase, even more preferably an alpha-1,3-fucosyltransferase, and as such plays a role in fertilization, embryogenesis, lymphocyte trafficking, immune response, cancer metastasis and resistance to various microorganisms. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 284 from positions 70 to 406, and 315 to 345. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 284 having any of the biological activity described herein. The glycosyltransferase activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in (Palcic et al. Carbohydr Res 1990;196:133-40).

Fucosylated compounds have considerable potential both as therapeutics and as reagents for clinical assays. However, synthesis of glycosylated compounds of potential commercial and/or therapeutic interest is difficult because of the very nature of the saccharide subunits. A multitude of positional isomers in which different substituent groups on the sugars become involved in bond formation, along with the potential formation of different anomeric forms, are possible. As a result of these problems, large scale chemical synthesis of most carbohydrates is not possible due to economic considerations arising from the poor yields of desired products. Enzymatic synthesis using glycosyl transferases such as fucosyltransferase provides an alternative to chemical synthesis of carbohydrates. Enzymatic synthesis using glycosidases, glycosyl transferases, or combinations thereof, have been considered as a possible approach to the synthesis of carbohydrates. As a matter of fact, enzyme-mediated catalytic synthesis would offer dramatic advantages over the classical

synthetic organic pathways, producing very high yields of carbohydrates economically, under mild conditions in aqueous solutions, and without generating notable amounts of undesired side products. To date, such enzymes are however difficult to isolate, especially from eukaryotic, e.g., mammalian sources, because these proteins are only found in low concentrations, and tend to be membrane-bound. In addition to being difficult to isolate, the acceptor (peptide) specificity of glycosyl transferases is poorly understood. Thus, there is a need for obtaining recombinant glycosyl transferase, including fucosyltransferases, that could be produced in very large amounts.

Thus, the invention related to methods and compositions using the protein of the invention or part thereof to synthesize glycosylated compounds, either glycoproteins, glycolipids, or oligosaccharides, more particularly fucosylated compounds. If necessary, the protein of the invention or part thereof may be produced in a soluble form by removing its transmembrane domains and/or its Golgi retention signal using any of the methods skilled in the art including those described in US patent 5,776,772. For example, the protein of the invention or part thereof is added to a sample containing GDP-fucose and a substrate compound in conditions allowing glycosylation, more particularly fucosylation and allowed to catalyze the glycosylation of this compound. In a preferred embodiment, the enzymatic reaction carried out by the protein of the invention is part of a series of other chemical and/or enzymatic reactions aiming at the synthesis of complex glycosylated compounds, such as the ones described in US patents 5,409,817 and 5,374,541. In another preferred embodiment where the method is to be practiced on a commercial scale, it may be advantageous to immobilize the glycosyltransferase on a support. This immobilization facilitates the removal of the enzyme from the batch of product and subsequent reuse of the enzyme. Immobilization of glycosyltransferases can be accomplished, for example, by removing from the transferase its membrane-binding domain, and attaching in its place a cellulose-binding domain. One of skill in the art will understand that other methods of immobilization could also be used and are described in the available literature.

In a preferred embodiment embodiment, the present invention relates to processes and compositions for producing glycosylated compounds, preferably fucosylated compounds, wherein a cell is genetically engineered to produce the protein of the invention or part thereof and used in combination with one or several other cells able to produce the donor substrate for the protein of the invention.

In another preferred embodiment, the present invention relates to a process and compositions for controlling the glycosylation of proteins in a cell wherein an insect, plant, or animal cell is genetically engineered to produce one or more enzymes that provide internal control of the cell's glycosylation mechanism. Preferably, the invention relates to a Chinese hamster ovary (CHO) cell line that is genetically engineered to produce a fucosyltransferase of the present invention either alone or in combination with other glycosyltransferases. This supplemental fucosyltransferase modifies the glycosylation machinery to produce glycoproteins having

carbohydrate structures that more closely resemble naturally occurring human glycoproteins. The methods for performing the above process and making the above compositions are carried out using the methods known in the art and described in U.S. Patent No. 5,047,335.

Another embodiment of the present invention relates to compositions and methods using the protein or part thereof to detect fucosylated conjugates. In a preferred embodiment, the protein of SEQ ID No: 284 or part of thereof is used to obtain reagents, such as antibodies. These reagents could be used in radioimmunoassays, competitive binding assays, Western Blot analysis, enzyme-linked immunosorbant assay (ELISA), immunohistochemisty, or any other technique known to those skilled in the art (Palcic et al. Carbohydr Res 1990;196:133-40). In a preferred embodiment, antibodies raised against the present protein of invention provides tools to specifically visualize salivary or digestive tract tissues (and cells derived from the tissues). This can be useful for various applications, including the determination of the origin or identity of cells, e.g. cancerous cells, as well as to facilitate the identification of particular cells and tissues for, e.g. the evaluation of histological slides. Such assays may also be used for diagnosis in various disorders including, but are not limited to, neoplastic tumors such as salivary, prostate, liver, digestive disease tract and pancreas cancers. Various types of samples can be assayed, including tumor tissues, or other biological samples such as serum or plasma.

The invention further relates to glycosylated compounds, preferably fucosylated compounds, obtained using any of the processes described herein using the protein of the invention or part thereof. Such compounds may be used in the diagnosing, prevention and/or treating of disorders including, but are not limited to, cancer, cystic fibrosis, ulcer, inflammation and immune based disorders, including autoimmune disorders such as arthritis, fertility disorders, and hypothyroidism. These conditions include infectious diseases where active infection exists at any body site, such as meningitis and salpingitis; complications of infections including septic shock, disseminated intravascular coagulation, and/or adult respiratory distress syndrome; acute or chronic inflammation due to antigen, antibody and/or complement deposition; inflammatory conditions including arthritis, cholangitis, colitis, encephalitis, endocarditis, glomerulonephritis, hepatitis, myocarditis, pancreatitis, pericarditis, reperfusion injury and vasculitis. Immune-based diseases include but are not limited to conditions involving T-cells and/or macrophages such as acute and delayed hypersensitivity, graft rejection, and graft-versus-host disease; auto-immune diseases including type I diabetes mellitus and multiple sclerosis. In a preferred embodiment, these glycosylated compounds or derivatives thereof may be used as pharmacological agents to trap pathogens or endogenous ligands thus reducing the binding of pathogens or endogenous ligands to the endogenous glycosylated compounds. For example, such compounds may be used to prevent and/or inhibit the adhesion of cancer cells to inner wall of blood vessel or aggregation between cancer cells and platelets, thus reducing cancer metastasis, to prevent and/or inhibit the adhesion of neutrophils to blood vessels endothelial cells. Other disorders include infections in which

recognition of a glycosylated product is essential to the development of the infection. Such infections include, but are not limited to those caused by Helicobacter pylori, E. coli and viruses such as HIV. In a preferred embodiment, such compounds, preferably oligosaccharides, are used as gram positive antibiotics and disinfectants (U.S. Pat. Nos. 4,851,338 and 4,665,060).

The invention further relates to methods and compositions using the protein of the invention or part thereof for diagnosis, prevention and/or treatment of several disorders in which recognition of glycosylated compounds, preferably of fucosylated compounds, is impaired or needs to be impaired. For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. For prevention and/or treatment purposes, inhibiting the endogenous expression of the protein of the invention may be used to reduce the production of glycosylated compounds detrimental to the organism using any of the antisense or triple helix methods described herein as well as antagonists of the protein's activity.

In another embodiment, various substances can be used for treatment, attenuation and/or prevention for treatment of abnormal conditions associated to unbalanced amounts and/or activity of the protein of SEQ ID No. 284. Such substances include, but are not limited to, chemical compounds such as agonists and antagonists, nucleic acids, and antibodies. In particular, the protein of the invention or part thereof may be used in the development of inhibitors of glycosyl transferase, more particularly inhibitors of fucosyltransferases, for mechanistic and clinical applications (Taylor, Curr Opin Struc Biol 1996;6:830-7 ; Colman, Pure Appl Chem 1995;67:1683-8; Bamford, Enz Inhib 1995;10:1-16 ; Khan & Matta, *In* Glycoconjugates, Composition, Structure, and Function. pp361-378. eds., Allen, H. J. & Kisailus, E. C. Marcel Dekker, Inc. New York, 1992 ; Thorne-Tjomsland et al., Transplantation 2000;69:806-8 ; Basset et al., Scand J Immunol 2000;51:307-11). Such substances may be employed for treatment of a variety of therapeutic and prophylactic purposes including certain types of neoplastic disorders. For instance, substances targeted against the protein of SEQ ID No. 284 can be administered to treat patients affected with, but not limited to, salivary, prostate, liver, digestive disease tract and pancreas cancers. Alternatively, such substances can be used for treatment, attenuation and/or prevention of infectious disease in order to induce resistance of various microorganisms agents.

Protein of SEQ ID NO:292 (internal designation 181-10-1-0-D10-CS)

The protein of SEQ ID NO:292 is encoded by the cDNA of SEQ ID NO:51. Accordingly, it will be appreciated that all characteristics and uses of the polypeptide of SEQ ID NO:292 described throughout the present application also pertain to the polypeptide encoded by a nucleic acid included in clone 181-10-1-0-D10-CS. In addition, it will be appreciated that all characteristics and uses of the nucleic acid of SEQ ID NO:51 described throughout the present application also pertain to the nucleic acid included in clone 181-10-1-0-D10-CS.

The protein of SEQ ID NO:292 was identified among the cDNAs from a library constructed from fetal liver. Tissue distribution analysis using databases indicated that mRNA encoding this protein was found primarily in fetal kidney and fetal liver.

The protein of SEQ ID NO:292 is most likely a polymorphic variant (92% identity) of human secreted protein SEQ ID NO: 197 from the protein described in PCT publication WO 9906553-A2, the disclosure of which is incorporated herein by reference in its entirety. Further, the protein of SEQ ID NO:292 is homologous to the C-type lectin domain of mouse macrophage asialoglycoprotein-binding protein (M-ASGP-BP, 36% identity), mouse natural killer (NK) cell surface protein P1 40 (NKR-, P1.9, 34%) and human asialoglycoprotein receptor L-H2 from EP 773289-A2 (27%). Thus, the present invention relates to nucleic acid and amino acid sequences of a lectin-like protein and to the use of these sequences in the diagnosis, study, prevention and treatment of disease.

The protein of SEQ ID NO:292 consists of 111 amino acids. From the amino acid alignments and the hydrophobicity plots, it has a predicted signal peptide sequence spanning residues 12-24 and one predicted transmembrane domain spanning residues 5-25. Accordingly, one embodiment of the present invention is a polypeptide comprising the signal peptide or the transmembrane domain.

A number of different protein families share a conserved domain which was first characterized in some animal lectins and which seems to function as a calcium-dependent carbohydrate-recognition domain (Drickamer K., J. Biol. Chem., 263:9557-9560,1988, the disclosure of which is incorporated herein by reference in its entirety). This domain, which is known as the C-type lectin domain (CTL) or as the carbohydrate-recognition domain (CRD), consists of about 110-130 residues. There are four cysteines that are perfectly conserved and involved in two disulfide bonds. Several categories of proteins can de found in which the CTL domain has been described. Both M-ASGP-BP and NKR-P1 are type II membrane proteins. Type II membrane proteins in which the CTL domain has been located at the C-terminal extremity include: 1) Asialoglycoprotein receptors (ASGPR), also known as hepatic lectins. The ASGPR's mediate the endocytosis of plasma glycoproteins to which the C-terminal sialic acid residue in their carbohydrate moieties has been removed. 2) A number of proteins expressed on the surface of NK cells, and some subsets of T cells: NKG2, NKR-P1, Ly-49, CD69, and on B cells: CD72, LyB-2. The CTL- domain in these proteins is distantly related to other CTL-domains, and it is unclear whether they all bind carbohydrates.

M-ASGP-BP is a lectin-like molecule expressed on the surface of activated macrophages and specific for terminal D-galactose and N-acetyl-D-galactosamine units (Oda S et al., J. Biochem., 104:600-605,1988, the disclosure of which is incorporated herein by reference in its entirety). Experimental results suggest that M-ASGP-BP participates in the interaction between tumoricidal macrophages and tumor cells.

ASGPR is a membrane protein expressed specifically by hepatocytes. Its function is to uptake asialoglycoproteins in the serum for degradation in the liver. Partially deglycosylated plasma glycoproteins are efficiently and specifically removed from the circulation by a receptor-mediated process. In mammals, the ASGPR specific for desialylated (galactosyl-terminal) glycoproteins, is expressed exclusively in hepatic parenchymal cells. Following binding of the ligand to this cell surface receptor, the receptor-ligand complex is internalized and transported by a series of membrane vesicles and tubules to an acidic-sorting organelle where receptor and ligand dissociate (Spiess M et al., J. Biol. Chem., 260:1979-1982, 1985, the disclosure of which is incorporated herein by reference in its entirety). Reduction in expression of AGPR has been reported in response to such liver conditions as hepatic cirrhosis, liver cancer and regenerated liver (Stadalnik et al., J. Nucl. Med., 26:1233-1242, 1985, the disclosure of which is incorporated herein by reference in its entirety). It has also been reported that ASGPR itself is present in serum (Katsugi et al., Alcohol Metabolism and Liver, 12:65-68, 1992, the disclosure of which is incorporated herein by reference in its entirety), which resulted in significant research being pursued toward the measurement of serum ASGPR. Furthermore, published results indicate that labeling compounds binding to ASGPR can be used as good indicators of liver function (Kudo, et al., Japan Assoc. of Gastrointest. Pathology., 89:1349-1359, 1992, the disclosure of which is incorporated herein by reference in its entirety).

NK cells constitute the third major population of lymphocytes. They possess the inherent capacity to kill various tumors and virally infected cells and mediate the rejection of allografts. These properties allow NK cells to have a major role in the regulation of innate immune responses in particular, and immunological functions in general. Members of the NKR-P1 family are type-II transmembrane C-type lectin receptors found on the surface of NK cells and a subset of T lymphocytes (NK T cells). Further, a subset of NKR-P1 molecules has been identified at the surface of peripheral blood monocytes and dendritic cells (Poggi A et al., Eur. J. Immunol., 27: 2965-2970, 1997, the disclosure of which is incorporated herein by reference in its entirety). Deficiencies in NKR-P1$^+$ T cells, which preferentially accumulate in the liver and bone marrow, have been implicated in the susceptibility to many diseases including insulin-dependent diabetes mellitus (IDDM, Tori M et al. Transplantation, 70:32-38, 2000, the disclosure of which is incorporated herein by reference in its entirety) and multiple sclerosis (Poggi A et al., J. Immunol., 162: 4349-4354, 1999, the disclosure of which is incorporated herein by reference in its entirety). NKR-P1 receptors have been shown to activate NK cell cytotoxicity coupled with release of interferon-γ (IFN-γ, Brown MG., Immunol. Rev., 155: 55-75, 1997, the disclosure of which is incorporated herein by reference in its entirety). However, unlike the well-characterized MHC class I ligands that regulate the specificity of the Ly-49 family of molecules, which are structurally related to the NKR-P1 receptors, cognate ligands for the NKR-P1 molecules have yet to be identified. Interestingly, it has been reported that a subset of the NKR-P1 molecules- NKR-P1B -

inhibits NK cell activation (Carlyle et al., J. Immunol., 162:5917-5923, 1999, the disclosure of which is incorporated herein by reference in its entirety).

Based on the structural and chemical homologies the protein of SEQ ID NO:292 was characterized as a C-type lectin-like, type II membrane protein, whose ligand binding may be calcium dependent. The protein of SEQ ID NO:292 or fragments thereof may provide the basis for clinical diagnosis of diseases associated with its induction and/or repression. This protein, framents thereof or antagonists/inhibitors thereof may be useful in the diagnosis and treatment of tumors, viral infections, inflammation, or conditions associated with impaired immunity, organ transplantation, bacterial infections, autoimmunity, hepatic dysfunction and liver regeneration. Furthermore, the protein SEQ ID NO:292 or fragments thereof may be used as a reagent for analyzing the control of gene expression by IFNs and other cytokines such as IL-12 and IL-4, as well as growth and transcription factors, in normal and diseased cells.

The protein of SEQ ID NO:292 has homology to the CTL domains of the ASPRG, M-ASGP-BP and NKR-P1 molecules. The protein of SEQ ID NO:292, in membrane-bound or soluble forms may have cytokine receptor activity, cell proliferation/differentiation activity, T cell activation activity, tissue growth regulating activity, receptor/ligand activity, signal transduction activity, to promote transendothelial migration, anti-inflammatory activity, tumor inhibition activity, among others. Accordingly, the protein SEQ ID NO:292 or fragments thereof may be used in diagnosis and treatment of diseases such as, but not limited to, autoimmune disorders such as autoimmune hepatitis, rheumatoid arthritis, Graves disease, systemic lupus erythematosus, Wegener's granulomatosis, sarcoidosis, polyarthritis, pemphigus, pemphigoid, erythema multiform, Sjogren's syndrome, inflammatory bowel disease, autoimmune encephalitis, myasthenia gravis keratitis, scleritis, Lupus Nephritis, and allergic encephalomyelitis; proliferative disorders including various forms of cancer such as leukemias, lymphomas (Hodgkins and non-Hodgkins), sarcomas, melanomas, adenomas, carcinomas of solid tissue, hypoxic tumors, squamous cell carcinomas of the mouth, throat, larynx, and lung, genitourinary cancers such as cervical and bladder cancer, hematopoietic cancers, head and neck cancers, and nervous system cancers, benign lesions such as papillomas, atherosclerosis, angiogenesis; viral infections, in particular HBV, HCV and HIV infections, as well as other viral- and pathogen-induced infections. The protein of SEQ ID NO:292 or fragments thereof may also be used to treat conditions associated with inflammation or immune impairment (e. g. reumathoid and osteo arthritis and AIDS), allergy, hepatic cirrhosis and liver toxicity; as well as genetic disorders, chronic illnesses and infections associated with decrease in NK, NK T, moacrophage, monocyte and dendritic cell functions. In another embodiment of the invention, inhibitors of the protein of SEQ ID NO:292 may be used to treat conditions such as multiple sclerosis, IDMM, graft versus host disease (GVH) and transplanted organ rejection.

Another embodiment relates to methods to treat and/or prevent the bacterial infections that arise in liver due to bacterial antigens brought from the intestine from the portal vein. In this

embodiment, the protein of SEQ ID NO:292 may be used to counteract the effects of the bacterial endotoxin lipopolysaccharide (LPS). Another embodiment of the invention is the use of the protein of SEQ ID NO:292 or fragments thereof to inhibit of NK cells activated by bacterial superantigens or LPS, which would help treat vascular endothelial injury in conditions such as Kwasaki disease.

The appearance of autoantibodies against the protein of SEQ ID NO:292 can be used as an indicator for autoimmune hepatitis (AIH), a disease that can lead to cirrhosis and fatal intractable hepatitis, as well as primary biliary cirrosis. The nucleic acid sequences encoding the protein of SEQ ID NO:292 or fragments thereof can be used for producing secreted forms of the protein. They can also be used to develop products for diagnosis and therapy. Accordingly, recombinant soluble derivatives can be used for detecting and measuring antibodies specific for the protein of the invention, e.g. by ELISA, Western blotting, etc. This allows AIH to be diagnosed and distinguished from other diseases.

In another embodiment of the invention, the protein of SEQ ID NO:292 or fragments or derivatives thereof can also be used for the analysis and purification of asialoglycoproteins and to develop inhibiting agents against asialoglycoprotein incorporation, or viral and other protein invasion, into liver cells.

Another embodiment of the present invention relates to polypeptides comprising the protein of SEQ ID NO:292 or fragments thereof and polynucleotides encoding the protein of SEQ ID NO:292 or fragments thereof. In another aspect the protein of SEQ ID NO:292 or fragments thereof may be used to identify specific molecules with which it binds such as agonists, antagonists or inhibitors. In a further aspect, the invention relates to methods for identifying agonists and antagonists/inhibitors of the protein of SEQ ID NO:292, and treating conditions associated with the protein of the invention or imbalance with the identified compounds. In a still further aspect, the invention relates to diagnostic assays for detecting diseases associated with inappropriate levels or activity of the protein of SEQ ID NO:292. Another embodiment of the invention relates to methods of measuring the amount of the protein of SEQ ID NO:292 in serum. Another embodiment relates to the use of labeling compounds that bind to the protein of SEQ ID NO:292 and can be used as good indicators of liver function or NK cell activity, among others.

An embodiment of the present invention relates to methods of using the protein of the invention or part thereof to identify and/or quantify or other ligands, which may interact with the protein of SEQ ID NO:292. The protein of SEQ ID NO:292 or fragments thereof may be include in pharmaceutical preparations for treating cancer or prevention/treatment of other diseases associated with changes in expression of the protein of the invention (see above). In a preferred embodiment of the invention the protein of the invention or part thereof is used to modulate the effect of cytokines and related molecule such as IL-1, IL-2, IL-12, IFN-γ. The protein of SEQ ID NO:292 may also be used to correct defects in vivo models of disease such as autoimmune, inflammation, pathogen-mediated infection, liver toxicity, allograft rejection, GVH, as well as tumor models, by

injecting the protein either intraperitoneally, intravenously, subcutaneously or directly in the diseased tissue.

The DNA encoding the protein of SEQ ID NO:292 or fragments thereof may be used in diagnostic assays for conditions/diseases associated with up-regulation or down-regulation of the expression of the protein of the invention (see above). The diagnostic assay is useful to distinguish between absence, presence, and excess expression of the protein and to monitor regulation of levels of the protein of during therapeutic intervention. The DNA may also be incorporated into effective eukaryotic expression vectors and directly targeted to a specific tissue, organ, or cell population for use in gene therapy to treat the above mentioned conditions, including tumors and/or to correct disease- or genetic-induced defects in any of the above mentioned proteins including the protein of the invention. The DNA may also be used to design antisense sequences and ribozymes, which can be administered to modify gene expression in NK, NK T, macrophages, monocytes and dendritic cells and to influence expression of cytokines such as IL-1, IL-2, IL-4, IL-12, and IFN-γ. In vivo delivery of genetic constructs into subjects can be developed to the point of targeting specific cell types, such as tumor where expression of the protein of SEQ ID NO:292 may be affected or is modulating the expression and/or activity of other proteins such as cytokines, growth factors, their receptors and/or tumor antigens. The DNA may also be used to identify unknown upstream sequences (e. g. promoters and regulatory elements) by standard techniques and for research into the control of gene expression by IFNs and other cytokines, as well as growth and transcription factors in normal and diseased cells. Hybridization probes are useful to detect DNA encoding the protein of SEQ ID NO:292 (or closely related molecules) in biological samples, and for mapping the naturally occurring genomic sequence to a particular chromosome/chromosome region. The DNA may be used to generate and/or treat in vivo animal models of disease, including susceptibility or resistance to infection, tumors, autoimmune conditions, GVH, allograft rejection and liver toxicity, based on vaccine, knock-out and transgene technologies.

Antibodies against the protein of SEQ ID NO:292 are useful for the diagnosis of conditions and disease associated with its expression and to quantify the protein of the invention (e. g. in assays to monitor patients during therapeutic intervention). Antibodies specific for the protein may include, but are not limited to, polyclonal, monoclonal, chimeric, single chain, Fab fragments produced by a Fab expression library. Neutralizing antibodies are especially preferred for diagnostics and therapeutics. Diagnostic assays for the protein of the invention include methods utilizing the antibody and a label to detect the protein of SEQ ID NO:292 in human body fluids or extracts of cells or tissues as well as methods for detecting or measuring antibodies against the protein of SEQ ID NO:292.

The protein of SEQ ID NO:292 and its catalytic or immunogenic fragments or oligopeptides thereof, can be used for screening therapeutic compounds in any variety of drug screening techniques including high throughput. Methods which may be used to quantitate the expression of

the nucleotide or protein of the invention include, but are not limited to, polymerase chain reaction (PCR), RT-PCR, RNAse protection, Northern blotting, enzyme-linked immunosorbent asay (ELISA), radioimmunoassay (RIA), fluorescent activated cell sorting (FACS), immunoprecipitation , and chromatography.

Accordingly, the protein of SEQ ID NO:292 or fragments thereof may be used to purify or enrich proteins containing carbohydrates. In such embodiments, the lectin of the present invention is placed in contact with carbohydrate-containing proteins under conditions which facilitate specific binding. The lectin of the present invention may be fixed to a solid support. After binding, specifically bound proteins are dissociated using appropriate salt or other conditions.

The protein of SEQ ID NO:292 or fragments thereof may also be used to regulate any of the activities described above, including the interaction between tumoricidal macrophages and tumor cells, the activity of NK cells, the treatment of bacterial infections resulting from bacterial antigens brought from the intestine, or to counteract the effects of bacterial LPS.

Accordingly, the present invention includes the use of the protein of SEQ ID NO:292 , fragments comprising at least 5, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, or 200 consecutive amino acids thereof, or fragments having a desired biological activity to treat or ameliorate a condition in an individual. For example, the condition may be any of those described above or an abnormality in any of the functions listed above. In such embodiments, the protein of SEQ ID NO:292, or a fragment thereof, is administered to an individual in whom it is desired to increase or decrease any of the activities of the protein of SEQ ID NO:292. The protein of SEQ ID NO:292 or fragment thereof may be administered directly to the individual or, alternatively, a nucleic acid encoding the protein of SEQ ID NO:292 or a fragment thereof may be administered to the individual. Alternatively, an agent which increases the activity of the protein of SEQ ID NO:292 may be administered to the individual. Such agents may be identified by contacting the protein of SEQ ID NO:292 or a cell or preparation containing the protein of SEQ ID NO:292 with a test agent and assaying whether the test agent increases the activity of the protein. For example, the test agent may be a chemical compound or a polypeptide or peptide.

Alternatively, the activity of the protein of SEQ ID NO:292 may be decreased by administering an agent which interferes with such activity to an individual. Agents which interfere with the activity of the protein of SEQ ID NO:292 may be identified by contacting the protein of SEQ ID NO:292 or a cell or preparation containing the protein of SEQ ID NO:292 with a test agent and assaying whether the test agent decreases the activity of the protein. For example, the agent may be a chemical compound, a polypeptide or peptide, an antibody, or a nucleic acid such as an antisense nucleic acid or a triple helix-forming nucleic acid.

In one embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify the source of a sample as, for example, fetal liver or fetal kidney, or to distinguish between two or more possible sources of a

sample on the basis of the level of the protein of SEQ ID NO:292 in the sample. For example, the protein of SEQ ID NO:292 or fragments thereof may be used to generate antibodies using any techniques known to those skilled in the art, including those described therein. Such antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry. In such methods a sample is contacted with the antibody, which may be detectably labeled, under conditions which facilitate antibody binding. The level of antibody binding to the test sample is measured and compared to the level of binding to control cells from fetal liver or fetal kidney or tissues other than fetal liver or fetal kidney to determine whether the test sample is from fetal liver or fetal kidney. Alternatively, the level of the protein of SEQ ID NO:292 in a test sample may be measured by determining the level of RNA encoding the protein of SEQ ID NO:292 in the test sample. RNA levels may be measured using nucleic acid arrays or using techniques such as in situ hybridization, Northern blots, dot blots or other technques familiar to those skilled in the art. If desired, an amplification reaction, such as a PCR reaction, may be performed on the nucleic acid sample prior to analysis. The level of RNA in the test sample is compared to RNA levels in control cells from fetal liver or fetal kidney or tissues other than fetal liver or fetal kidney to determine whether the test sample is from fetal liver or fetal kidney.

In another embodiment, antibodies to the protein of the invention or part thereof may be used for detection, enrichment, or purification of cells expressing the protein of SEQ ID NO:292, including using methods known to those skilled in the art. For example, an antibody against the protein of SEQ ID NO:292 or a fragment thereof may be fixed to a solid support, such as a chromatograpy matrix. A prepartation containing cells expressing the protein of SEQ ID NO:292 is placed in contact with the antibody under conditions which facilitate binding to the antibody. The support is washed and then the cells are released from the support by contacting the support with agents which cause the cells to dissociate from the antibody.

In another embodiment of the present invention, the protein of SEQ ID NO:292 or a fragment thereof thereof may be used to diagnose disorders associated with altered expression of the protein of SEQ ID NO:292. In such techniques, the level of the protein of SEQ ID NO:292 in an ill individual is measured using techniques such as those described herein. The level of the protein of SEQ ID NO:292 in the ill individual is compared to the level in normal individuals to determine whether the individual has a level of the protein of SEQ ID NO: 292 which is associated with disease.

Protein of SEQ ID NO: 408 (internal designation 179-14-2-0-F11-CS

The 236 amino acid protein of SEQ ID NO: 409, herein referred to as PNMT A, and encoded by the cDNA of SEQ ID NO: 168 is found in fetal kidney and fetal brain. PNMT A is a

polymorphic variant of human phosphotidylethanolamine N-methyltransferase (PNMT) (SPTREMBLNEW SPTREMBL SWISSPROT accession number Q9UHY6). PNMT A differs from the sequence of PNMT (STR accession number Q9UHY6) by two amino residues. Position 95 contains an isoleucine residue (I) substituted for a valine residue (V); position 130 contains a valine residue (V) substituted for a glutamine residue (G). PNMT A displays 4 candidate membrane-spanning segments in positions 50 to 70, 83 to 103, 131 to 151 and 196 to 216.

Catecholamine neurotransmitters [*e.g.*, dopamine, noradrenaline (norepinephrine), adrenaline (epinephrine)] are synthesized in catecholaminergic neurons from tyrosine, via dopa, dopamine and noradrenaline, to adrenaline. Four enzymes are involved in the biosynthesis of adrenaline: (1) tyrosine 3-mono-oxygenase (tyrosine hydroxylase, TH); (2) aromatic L-amino acid decarboxylase (AADC, or Dopa decarboxylase, DCC); (3) dopamine beta-mono-oxygenase (dopamine beta-hydroxylase, DBH); and (4) noradrenaline N-methyltransferase (phenylethanolamine N-methyltransferase, PNMT)(Nagatsu, Neurosci. Res. 12:315-345 (1991)). PNMT, the final enzyme in the pathway for adrenaline biosynthesis catalyses the production of adrenaline from noradrenaline using S-adenosyl-L-methionine as a methyl donor. For this reason, PNMT serves as a good marker for tissues and cells producing epinephrine (adrenaline). Studies conducted by Kennedy and collaborators have shown that PNMT are widely distributed in human tissues including heart and kidney (Kennedy et al., J. Clin. Invest. 95:2896-2902 (1995)).

In some pheochromocytomas, the tumors contain and secrete greater amounts of adrenaline than do normal adrenal medullas. In a case/control study, Isobe *et al.* have shown that adrenaline-secreting pheochromocytomas express significantly greater amounts of PNMT mRNA than do normal adrenal medullas (Isobe *et al.* J. Urol. 163:357-362 (2000)). Moreover, PNMT immunoreactivity is only detected in the adrenaline-secreting tumors. The C-1 region in the rostral ventral lateral medulla contains mainly adrenaline neurons. These neurons are the tonic vasomotor center of the brain. Burke *et al.* have demonstrated changes in the enzymatic activity of PNMT in axon terminals and cell bodies of neurons from the medulla of patients with Alzheimer's disease. They have also shown that PNMT protein is decreased in axon terminals in brains from patients with Alzheimer's disease; the decrease in PNMT appears to be due to retrograde degeneration of epinephrine neurons (Burke *et al.*, Ann. Neurol. 22:278-280 (1987)). In the case of advanced Alzheimer's disease, the Burke et al. presented evidence that the accumulation of PNMT in the perikarya results from diminished transport of this enzyme to axon terminals (Burke et *al.*, J. Am. Geriatr. Soc. 38:1275-1282 (1990)).

Neurons that contain PNMT have cell bodies in brain stem regions of the rat brain and send projections mainly into other brain stem areas, such as the hypothalamus and the spinal cord. These neurons can be affected pharmacologically by various kinds of drugs. PNMT inhibitors currently represent the only means of modifying adrenaline neurons pharmacologically without affecting noradrenaline or dopamine neurons in brain. Experiments conducted in deoxycorticosterone

acetate-salt (DOCA-salt) hypertensive rats and spontaneously hypertensive rats (SHR) have shown that inhibitors of PNMT lower blood pressure (Goldstein *et al.*, Life Sci. 30:1951-1957 (1982); Lyang *et al.*, Res. Commun. Chem. Pathol. Pharmacol. 46:319-329 (1984); Chatelain *et al.*, J. Pharmacol. Exp. Ther. 252:117-125 (1990)). Molecules and compounds affecting adrenaline neurons may also be of use in the treatment of psychiatric disorders and neuroendocrine dysfunction.

One embodiment of the subject invention provides polypeptides comprising the sequence of PNMT A. Other polypeptides of the invention include polypeptides comprising the amino acids of SEQ ID NO: 409 from positions 50 to 70, 83 to 103, 131 to 151 and/or 196 to 216. Also encompassed by the instant invention are biologically active fragments of the PNMT A protein. "Biologically active fragments" are defined as those peptide or polypeptide fragments of PNMT A which have at least one of the biological functions of the PNMT A protein (e.g., the ability to catalyze the formation of adrenaline). In a preferred embodiment, the biologically active fragment of PNMT A contains at least one of the amino acid substitutions which distinguish PNMT A from PNMT (*i.e.*, an isoleucine residue (I) substituted for a valine residue (V) at position 95; and/or valine residue (V) substituted for a glutamine residue (G) at position 130). In one embodiment, the PNMT A polypeptides of the invention are encoded by clone 179-1462-0-F11-CS.

Thus, one embodiment of the invention provides an enzymatic component of the adrenaline synthetic pathway and methods of producing adrenaline in accordance with methods known to those skilled in the art. These methods substitute PNMT A, or biologically active fragments thereof, for the PNMT enzyme used in these known synthetic pathways.

The invention also provides variants of the protein of SEQ ID NO: 409. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence of PMNT A. Variants according to the subject invention also have at least one functional or structural characteristic of PNMT A, such as the ability to catalyze the formation of adrenaline. The invention also provides biologically active fragments of the variant proteins. Unless otherwise indicated, the methods disclosed herein may be practiced utilizing PNMT A or variants thereof. Likewise, the methods of the subject invention may be practiced using biologically fragments of PNMT A, or variants thereof, provides that said biologically active fragments contain the amino acid substitutions noted supra.

One embodiment of the subject invention provides methods of using the protein of the invention, or biologically active fragments thereof, to label (chemically or isotopically) the adrenaline molecule *in vitro*. The labeled adrenaline molecules can then be used to localize receptors in tissue cuts by *in situ* hybridization experiments.

The invention also provides a fusion protein or polypeptide in which PNMT A, or biologically active fragments thereof, are combined with another protein (tag) by the use of a recombinant DNA molecule. The resulting purified, and enzymatically active fusion product, is then

added, *in vitro,* to the noradrenaline precursor and to S-adenosyl-L-methionine as a methyl donor. The enzymatic reaction is then performed in conditions known to those skilled in the art (Burke *et al.*, Proc. Soc. Exp. Biol. Med. 181:66-70 (1986); Morimoto *et al.*, Endocr. J. 40:179-183 (1993)). In this reaction, the methyl group of S-adenosyl-L-methionine must be labeled isotopically ([14C]-S-adenosyl-L-methionine or [methyl-3H]-S-adenosylmethionine), or chemically, in order to allow the transfer of a "tagged" methyl group to the adrenaline molecule.

Similarly, in cells transfected with cDNAs encoding the protein of the invention PNMT activity of expressed proteins may be measured by incubating cytosolic fractions with [14C]-S-adenosyl-L-methionine and normetanephrine for 60 min according to methods described by those skilled in the art (Morimoto *et al.*, ibid.). Agonists and/or antagonists of PNMT activity may also be tested (high throughput screening) on transfected cells expressing the wild type form of the protein of the invention. Again, effects of such drugs on PNMT enzymatic activity is measured by the methods described above.

The invention further relates to methods and compositions used to modify the protein of the invention (i.e. derivatize the PNMT A protein). Post-translational modifications encompassed by the invention include, N-linked or O-linked carbohydrate chains, processing of N-terminal or C-terminal ends, attachment of chemical moieties, such as polyethylene glycol, to the amino acid backbone, chemical modifications of N-linked or O-linked carbohydrate chains, and addition or deletion of an N-terminal methionine residue as a result of prokaryotic host cell expression. Some of these modifications of the protein of the invention may facilitate its extraction and purification in prokaryotic expression systems. Post-translational modifications such as N-linked or O-linked carbohydrate chains addition may also optimize the enzymatic activity of the protein of the invention when it is first produced in a prokaryotic system.

Another embodiment of the subject invention provides antibodies directed against the protein of the invention or immunogenic fragments thereof. The antibodies of the invention are useful for the screening of tissues and cells producing adrenaline or for affinity purification of PNMT or PNMT A. These antibodies may also be used in the diagnosis of pathologies and disorders such as pheochromocytomas and Alzheimer's disease, where PNMT A is overexpressed. Methods of peforming affinity purification as well as methods of making polyclonal and monoclonal antibodies are well known to those skilled in the art.

In therapeutic regimens, neutralizing antibodies may be used as antagonists of PNMT A and used to treat conditions associated with overexpression of PNMT A. These disorders include, and are not limited to, hypertension, pheochromocytomas, and advanced Alzheimer's disease (Goldstein *et al.*, Life Sci. 30:1951-1957 (1982); Lyang *et al.*, Res. Commun. Chem. Pathol. Pharmacol. 46:319-329 (1984); Chatelain *et al.*, J. Pharmacol. Exp. Ther. 252:117-125 (1990); Isobe et al., J. Urol. 163:357-362 (2000); Burke et *al.*, J. Am. Geriatr. Soc. 38:1275-1282 (1990)).

Proteins of SEQ ID NOs: 395 and 403 (internal designation: 160-101-3-0-H2-CS and 160-99-4-0-E4-CS respectively

The 367-amino-acid-long proteins of SEQ ID NOs: 395 and 403 encoded by the cDNAs of SEQ ID NOs: 154 and 162 respectively are polymorphic variants, the first one being overexpressed in fetal brain and ovary and the second one in fetal brain only. They both contain glutathine S-transferase (GST) domains from positions 47 to 122, and 206 to 309 which are respectively the G-site and H-site described below. In addition, they also display two hydrophobic domains (from aa 258 to aa 278 and from aa 338 to aa 358) which are characteristic of some GST proteins.

Glutathione S-transferase proteins (GSTs) are dimeric proteins that catalyse the conjugation of glutathione to a wide range of hydrophobic compounds (through the formation of a thioether bond with their electrophilic centre) to create the products which are less reactive, more hydrophilic, and thus more easily excreted from the cells. The GST superfamily (E.C. 2.5.1.18) is indeed believed to be one of the most important proteins in the detoxification of reactive electrophiles within living cells. Glutathione is a cellular tripeptide (gamma-glutamylcysteinylglycine) which is perhaps the most abundant amino acid derivative contained in the cells of higher life forms. The middle amino acid in glutathione, cysteine, has a free thiol group which can compete with the nucleophilic site on nucleotide bases for reaction with electrophiles. Within the cell, glutathione functions so as to conjugate to xenobiotic toxic molecules in general, and electrophiles in particular, to render the toxic molecules less reactive against cellular macromolecules and to target the toxic molecules for subsequent metabolic and excretion pathways.

Based on amino acid sequence identity, there are at least seven major classes of GST proteins (designated alpha, kappa, mu, pi, sigma, theta and zeta). Sequence similarity between classes is rather low, ranging between 20-30%. However, a single point mutation in the H-subsite region of GST is enough to shift substrate specificity from class pi to alpha (Nuccetelli M.N. et al. Biochem.Biophys.Res.Commun. 252: 184-189 (1998)). In spite of relatively low sequence identity, the GSTs exhibit a high degree of structural similarity. It is generally known that the GST molecule binds quite specifically and with high affinity to glutathione, but binds promiscuously to a wide variety of xenobiotic, electrophilic, and alkylating chemical agents. All GST enzymes of the four main cytosolic classes is found in dimeric form with two active sites per dimer each of which functions independently of the other. The active site has been characterized as consisting of a glutathione binding region (designated the G-site) and a non-specific hydrophobic binding region (designated the H-site) to accommodate the electrophilic substances. Pi-, mu-, alpha- and theta-class crystal structures have been elucidated; all possess a similar GSH-binding site, but the hydrophobic substrate-binding site (H-subsite) is subject to variation across the classes (Allardyce C.S. et al. Biochem.J. 343 525-531 (1999)). The GST activity has been suggested be involved in the regulation of the assembly of multisubunit complexes by shifting the balance between glutathione, disulfide glutathione, thiol groups of cysteines, and protein disulfide bonds. The GST domain is a

widespread, conserved enzymatic module that may be covalently or noncovalently complexed with other proteins. Regulation of protein assembly and folding may be one of the functions of GST (Koonin EV et al. Protein Sci 3:2045-2054 (1994)).

The cytosolic glutathione S-transferase are known to belong to four classes, designated Alpha, Mu, Pi and Theta. A fifth class of glutathione S-transferases is a microsomal enzyme found primarily in liver endoplasmic reticulum. An extensive analysis of the expression microsomal glutathione transferase 1 in human tissues shows that it predominantly occurs in liver and pancreas. The relative expression levels in man ranged from: liver and pancreas to kidney, prostate, colon (30-40%), heart, brain, lung, testis, ovary, small intestine (10-20%), placenta, skeletal muscle, spleen, thymus and peripheral blood leucocytes (1-10%). Liver-enriched expression was detected in human fetal tissues with lung and kidney displaying lower levels (10-20%). No transcripts could be detected in fetal brain or heart (Estonius M et al. Eur J Biochem 260:409-13 (1999)). Based on these observations, and the fact that the enzyme is encoded by a highly conserved single-copy gene, it is suggested that microsomal glutathione transferase 1 performs essential functions vital to most mammalian cell types. One particular glutathione S-transferase was still identified in mitochondrial matrix (Pemble S.E. et al. Biochem.J. 319 : 749-754 (1996)).

GST and GST-like proteins are largely spread in organisms. In vertebrates and in cephalopodes some proteins (christallins) presented in the lenses are structurally related to alpha-class GSTs (Chiou S.H. et al. Biochem.J. 309 : 793-800 (1995)). Furthermore, the olfactory epithelial cytosol shows the highest GST activity among the extrahepatic tissues. The olfactory GSTs were found to catalyse glutathione conjugation of several odorant classes, including many unsaturated aldehydes and ketones, as well as epoxides and were proposed to play an important rôle in chemoreception (Ben-Arie N. et al. Biochem.J. 292 : 379-384 (1993)).

Higher cells each contain a family of many GST isozymes in each class with broad, yet overlapping, specificity. Mu-class GSTs are thought to be involved in the detoxification of reactive oxygen species (cyclised o-quinones) produced via oxidative metabolism of catecholamines. These toxins are thought to be involved in neurological disorders of the nigrostriatal and mesolimbic systems (Parkinsons and Schizophrenia, respectively). Enzymes of the mu-class GSTs are expressed in the substantia nigra and have preferential substrate specificity for the cyclised o-quinones formed by catecholamine metabolism (Hansson L.O.et al. J.Mol.Biol. 287: 265-276 (1999), Takahashi Y.et al. J. Biol. Chem.268: 8893-8 (1993)). Whilst most of the GSTs share common substrates, there are distinct differences in substrate preference between subfamilies. These enzymes have evolved as a cellular protection system against a wide variety of electrophilic compounds, including a range of xenobiotics, oxidative metabolism by-products (oxidized lipid, DNA and catechols), and in particular are known to metabolise a number of environmental carcinogens.

GSTs are also known to catalyze other reactions, such as peroxidase and isomerase reactions (Edwards R. et al. Trends in Plant Sci. 5 : 193-198 (2000)) as well as the addition of

aliphatic epoxides and arene oxides to glutathione; the reduction of polyol nitrate by glutathione to polyol and nitrite; certain isomerization reactions and disulfide interchange. As well, there are marked species differences in catalytic activities between various purified mammalian hepatic GST mixtures. Some of them catalyse chemical stereospecific conversion of several pharmacological substances much less effective then anothers. For exemple, recombinant human GST was succesfully used in the reaction of steric conversion of 13-cis-retinoic acid to all-trans-retinoic acid (Chen H. and Juchau M.R. Biochem.J. 336 : 223-226 (1998)).

An increasing number of GST genes are being recognized as polymorphic. Certain alleles, particularly those that confer impaired catalytic activity may be associated with increased sensitivity to toxic compounds. Genetic polymorphisms and differences in GST expression have been implicated in individual susceptibility to certain types of cancer (for rev. Hayes JD and Strange RC Pharmacology 61:154-166 (2000). For exemple, GSTM1 deficiency predisposes to head and neck cancer, especially to cancer of the larynx, which is particularly exposed to tobacco smoke carcinogens (Gronau S et al. Laryngorhinootologie 79:341-344 (2000)). Conversely, over-expression of GSTs is thought to be involved in the phenomenon of multi-drug resistance to cancer chemotherapy. One of the class of electrophilic compounds that are substrates for the glutathione S-transferase enzymes is the group of alkylating agents used in antineoplastic therapy. A common problem that is observed in modern cancer chemotherapy is the appearance of chemotherapeutic resistant tumor cells that, because of the resistivity, no longer respond appropriately to the antineoplastic agents. This resistance is often observed with many drugs that have no physical or mechanistic similarities to the original agent. GST isoenzymes have been shown to be involved in the development of drug resistance to a variety of chemotherapeutic agents such as adriamycin, vinblastine, actinomycin D and colchicine (Beckett, et al. Adv. Clin. Chem. 30:281-380 (1993)). It has been demonstrated that a resistant population of malignant cells shows a modified pattern of total glutathione S-transferase activity. A resistant population of MCF-7 breast cancer cells, identified through selection in adriamycin by Batist et al., J. Biol. Chem., 261:15544-15549 (1986) resulted in a subset of cells which were approximately 200 fold more resistant than the parental cells. The resistant cells were found to exhibit a 45 fold increase in total glutathione S-transferase activity, the increase being due to the result of an appearance of an isozyme not expressed in the parental cell line. It was demonstrated that an increase in glutathione S-transferase alone, an increase conditioned by the transformation of susceptible cells with a foreign DNA construct expressing the wild-type glutathione S-transferase coding region, could increase the resistance of cells to an antineoplastic agent. As reported in Puchalski and Fahl, Proc. Natl. Acad. Sci. USA, 87:2443-2447 (1990), expression of the rat 1-1, 3-3 and the human P1-1 isozymes of glutathione S-transferase in COS cells increased their resistance to the agent. The recent study of increase in resistance of tumor cells to cytotoxic drugs or ionizing radiation has allowed to identify using differential display a new GST-related protein p28 expressed exclusively in lymphoma cell (Kodym

R. et al. J. Biol.Chem. 274: 511-5137 (1999)). Subcellular protein fractionation revealed p28 localization in the cytoplasm, but with thermal stress p28 relocated to the nuclear fraction of cellular proteins. The sequence homology and the similar functional characteristics of p28 to other GST family members (in particular relocalization in response to thermal stress and ability to bind glutathione), argues that p28 is a new mammalian member of GST superfamily.

Evidence suggests that the level of expression of GSTs is a crucial factor in determining the sensitivity of cells to a broad spectrum of toxic chemicals. In humans, marked interindividual differences exist in the expression of class alpha, mu and theta GST. For the most abundant mammalian classes of GST the mechanisms of transcriptional and post-translational regulation have been studied. The biological control of alpha-, mu- and pi- classes exhibit sex-, age-, tissue-, species-, and tumor-specific patterns of expression. In addition, GST are regulated by a structurally diverse range of xenobiotics and, to date, more then 100 chemicals have been identified that induce GST (Hayes J.D. and Pulford D.J. Crit.Rev.Biochem.Mol.Biol. 30 : 445-600 (1995)). A significant number of these chemicals occur naturally and, as they are found as nonnutrient components in vegetables and citrus fruits, it is apparent that humans are likely to be exposed regularly to such compounds. Many inducers effect transcriptional activation of GST genes through either the antioxidant-responsive element (ARE), the xenobiotic element (XRE), the GST P enhancer 1 (GPE), or the glucocorticoid-responsive element (GRE). Many of compounds that induce GST are themselves substrates for these enzymes, or are metabolized (by cytochrome P-450 monooxygenases) to compounds that can serve as GST substrates, suggesting that GST induction represent part of an adaptive response mechanism to chemical stress by electrophiles. It also appear probable that GST are regulated in vivo by reactive oxygen species, the potents inducers capable of generating free radicals by redox-cycling ; such relulation can be an adaptive response to oxydative stress in the cell. It has been shown GST-pi can potently and selectively inhibit activation of jun protein by its upstream kinase (JNK) ; these results suggest GST-pi can also be a regulator of signal transduction (Monaco R. et al. J.Prot.Chem. 18 : 859-866 (1999)). The majority of human tumors express significant amounts of class pi GST (Hayes&Pulford, supra).

Therefore, GSTs have medical importance due to their role in mediating drug resistance in cancer patients. The measurement of GST isoenzymes in vitro has importance in diagnostic medicine. For example, the measurement of the pi isoenzyme of GST in tissue specimens is useful in pathology for the detection and diagnosis of a variety of different tumors. In addition, measurement of the alpha form of GST in blood is useful for the detection and monitoring of a variety of different forms of liver disease (for a detailed description of the clinical applications of GST measurements see Beckett, et al., supra).

It is believed that the proteins of SEQ ID NOs: 395 and 403 or part thereof are transferases, probably transferring alkyl or acyl groups different from methyl group, more probably glutathione S-transferases and, as such, play a role in cellular detoxification especially against xenobiotics and

oxidative metabolism byproducts. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NOs: 395 and 403 from positions 47 to 122, and 260 to 309. Other preferred polypeptides of the invention are fragments of SEQ ID NOs: 395 and 403 having any of the biological activities described herein. The transferase activity of the proteins of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described for GST proteins as in US patents 5,866,792 and 6,096,504, which disclosures are hereby incorporated by reference in their entireties.

To find substrates, the proteins of the invention, or part thereof, or derivative thereof, may be used for screening libraries of compounds in any of a variety of drug screening techniques. The fragment employed in such screening may be free in solution, affixed to a solid support, borne on a cell surface, or located intracellularly. The formation of binding complexes, between the proteins of the invention, or part thereof, or derivative thereof, and the agent being tested, may be measured. Antagonists or inhibitors of the proteins of the invention may be produced using methods which are generally known in the art, including the screening of libraries of pharmaceutical agents to identify those which specifically bind the protein of the invention. Another technique for drug screening which may be used provides for high throughput screening of compounds having suitable binding affinity to the proteins of the invention as described in published PCT application WO84/03564.

The invention relates to methods and compositions using the proteins of the invention or part thereof or derivative thereof to catalyze GST-dependent detoxification reactions in vitro or in vivo using any methods known to those skilled in the art. For example, uses of the proteins of the invention or part thereof may be very useful to treat toxic byproducts such as the ones obtained in laboratory experiments, such as dietary toxins due to the use of pesticides on plants used to feed animals or humans, etc... Preferably, the proteins of the invention or part thereof or derivative thereof is added to a sample containing the substrate(s) in conditions allowing detoxyfication, and allowed to catalyze the detoxification of the substrate(s). In a preferred embodiment, the detoxification is carried out using a standard assay such as those described herein.

In some of the above cited embodiments, compositions comprising the proteins of the present invention or part thereof are added to samples as a "cocktail" with other detoxifying enzymes. The advantage of using a cocktail of detoxifying enzymes is that one is able to detoxify a wide range of substrates without knowing the specificity of any of the enzymes. Using a cocktail of detoxifying enzymes also protects a sample from a wide range of future unknown toxic compounds from a vast number of sources. For example, the proteins of the invention or part thereof is added to samples where toxic compounds are undesirable. Alternatively, the protein of the invention or part thereof may be bound to a chromatographic support, either alone or in combination with other detoxifying enzymes, using techniques well known in the art, to form an affinity chromatography column. A sample containing the undesirable substrate is run through the column to remove the substrate. Immobilizing the proteins of the invention or part thereof on a support is particularly

advantageous for those embodiments in which the method is to be practiced on a commercial scale. This immobilization facilitates the removal of the enzyme from the batch of product and subsequent reuse of the enzyme. Immobilization of the protein of the invention or part thereof can be accomplished, for example, by inserting a cellulose-binding domain in the protein. One of skill in the art will understand that other methods of immobilization could also be used and are described in the available literature. Alternatively, the same methods may be used to identify new substrates.

In a preferred embodiment, the invention relates to cells and plants or animals genetically engineered to express the protein of the invention or part thereof, preferably at a high level using any method known to those skilled in the art. Such engineered cells, animals or plants will display enhances detoxification of compounds. In a more preferred embodiment, expression of the proteins of the invention or part thereof will confer resistance to herbicides to transgenic plants using techniques similar to those described in the US patent 5,866,792.

For such embodiments, the proteins of the invention may need to be modified to enhance their ability to react with specific substrates. These modifications can provide novel isoforms which are specifically efficient against selected electrophilic or alkylating agents. Artificial DNA constructs encoding and expressing such modified or mutant proteins of the invention or part thereof may be selectively delivered into targeted cells to enhance the resistivity of those cells to the alkylating or neoplastic agents. The methods related to such modifications are described (Fahl , et al. United States Patent 6,136,605 Oct24, 2000). The method is based on random mutation and selection with the selection being performed with the agent against which enhanced activity is sought. The mutation is preferably site directed to the amino acids associated with the H-site on the enzyme, so as to favor the creation of new, useful isoforms of the enzyme.

In another embodiment, the invention relates to compositions and methods using the proteins of the invention or part thereof to design specific systems of artificial chemoreception as described in Ben-Arie N. et al, supra. Such chemoreception systems could recognize odorants, xenobiotics, pesticides, drugs and may be useful for chemical, cosmetic, pharmaceutical, forensic and any other analytical purposes. The design of such a system may be generally based on the subtle specificity of recognition of compounds by different GST isoenzymes. The methods to produce analytical diagnostics based on enzyme specificity are known by those skilled in the art.

In another embodiment, the invention relates to compositions and methods using the proteins of the invention or part thereof such as ligands for substrates of interest. In a preferred embodiment, the proteins of the invnetion or part thereof may be used to identify and/or quantify substrates using any techniques known to those skilled in the art such as those described in Koonin E., supra. In another preferred embodiment, the proteins of the invention or part thereof may be used to improve or to modify some molecular biology methods based on protein-protein interactions, including but not limited to two-hybrid assays, expression and purification systems

based on GST fusion to heterologous proteins as already available commercially (expression vectors or plasmids encoding fusions proteins and affinity purification methods).

In still another embodiment, the invention relates to methods and compositions using the proteins of the invention or part thereof as a marker protein to selectively identify tissues, preferably fetal brain for the protein of SEQ ID NO: 403, and preferably fetal brain and ovary for the protein of SEQ ID NO: 395. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

In another embodiment of the invention, measurement of the activity or expression of the proteins of the invention, may be used for the assessment of organ status, including organ damage following immunological or toxological insult and diagnostic of transplant rejection, using any technique known to those skilled in the art including those described in US patents 6,080,551 and RE35,419.

In another embodiment of the invention, the proteins of the invention or part thereof, or derivative thereof, may be used to diagnose, treat and/or prevent cell proliferative disorders linked to dysregulation of gene expression of the proteins of the invention. Such disorders include but are not limited to, benign tumors, and cancers such as adenocarcinoma; leukemia; melanoma; lymphoma; sarcoma; and cancers of the brain, ovary, bladder, colon, liver, small intestine, large intestine, breast, kidney, lung, and prostate. Diagnosis may be performed using nucleic acids or antibodies able to detect the expression of the protein of the invention using any technique known to those skilled in the art including Northern blotting, RT-PCR, immunoblotting methods immunohistochemisty, enzyme-linked immunosorbant assay (ELISA) described herein. Quantities of the protein of the invention expressed in subject samples, control and disease from biopsied tissues or body fluids or cell extracts taken from patients are compared with the standard values. Deviation between standard and subject values establishes the parameters for diagnosing disease.

For prevention and/or treatment purposes, the expression of the proteins of the invention may be enhanced using any methods known to those skilled in the art. For example, gene therapy techniques may be used such as the delivery of sense promoter polynucleotide constructs for the proteins of the invention or part thereof using a recombinant expression vector such as a chimeric virus or a colloidal dispersion system (see Nelson et al. United States Patent 552,277). Alternatively, the proteins of the invention or fragments thereof or derivatives thereof may be administered to a subject to treat or prevent cancerous and precancerous disorders as well as a proliferative disorders in general. Such disorders can include, but are not limited to, syndromes represented by abnormal neoplastic, including dysplastic, changes of tissue, dysplastic growths in ovary, brain, colonic, breast, prostate or lung tissues, dysplastic nevus syndromes, polyposis

syndromes, colonic polyps, precancerous lesions of the cervix (i.e., cervical dysplasia), esophagus, lung, prostatic dysplasia, prostatic intraneoplasia, breast and/or skin and related conditions (e.g., actinic keratosis), whether the lesions are clinically identifiable or not.

The invention also relates to compositions and methods using the proteins of the invention or part thereof or derivative thereof to decrease drug resistance in cancer chemotherapy. Inhibition of the expression and/or activity of the proteins of the invnetion may be achieved using any mean known to those skilled in the art. In a preferred embodiment,gene therapy methods such as antisense oligonucleotides, triple helices strategies are described elsewhere in the application. In another preferred embodiment, antagonists of the activity of the proteins of the invention may be used. These antagonists may be directly administered to patients. Low-molecular-weight inhibitors (i.e., those which can be delivered freely into the brain and which specifically inhibit GST activity) are especially preferred for use in cancer therapy. Alternatively, artificial DNA constructs encoding peptide modulators or inhibitors of the activity of the protein of the invention and flanking sequences effective to express the protein coding sequence in a host cell as well as flanking regulatory sequences (such as an antioxidant responsive element which enhances the expression of the glutathione S-transferase in the presence of antioxidant molecules) may be used. Such artificial DNA constructs confers to recombinant cells an increased level of resistance to an antineoplastic agent.

There is a need for selective inhibitors of GST isoenzymes for treatment of drug resistance in cancer patients. Thus, in a further embodiment of this invention, the proteins of the invention or fragments thereof can be used for screening of the compounds which are selective inhibitors or specific inhibitors of one or more GST isoenzymes. Selective inhibition means that a compound has a greater inhibitory effect on one isoenzyme than it does on another GST isoenzyme. Such compounds could also be tested and selected for their ability to overcome drug resistance to chemotherapeutic agents (see Jones , et al. United States Patent 6,103,665 Aug 2000). For example, mammalian cell lines that have been made resistant to particular chemotherapeutic drugs can be used to identify haloenol lactone compounds that render the lines sensitive to the chemotherapeutic agents. Such cell lines are known to those of skill in the art and can be obtained for example from the American Type Culture Collection, Rockville, Md., USA.

Protein of Seq Id No: GRP (187-38-0-0-110-CS)

The protein of SEQ ID No: 306, herein referred as GRP, encoded by the cDNA of SEQ ID No: 65 herein referred as GRP2, is homologous to bovine glutamic-acid rich protein (GARP) (GENEPEPT ID: M61185). The protein of the invention is overexpressed in the brain and fetal brain, lymph ganglia and thyroid.

The protein of the invention exhibits homology with bovine glutamic-acid rich protein (GARP) (18 % identical amino acids, 28% positive amino acids when aligned by BLASTP 2.0.9).

GARP proteins have been identified as multivalent proteins that interact the key players of cGMP signaling, phosphodiesterase and guanylate cyclase. GARP proteins are closely associated to cyclic nucleotide-gated channels (CNGs) which make up a family of nonselective cation channels found in a variety of tissues. The beta subunit of CNGs have a unique bipartite structure, containing a membrane-spanning region (beta part) and a GARP part (GARP). GARP is highly homologous to a soluble splice form, GARP1, and a splice variant lacking the C-terminal glutamic-acid-rich region. Experiments using GARP attached to affinity columns showed that phosphodiesterases are highly retained by the column [Korschen HG et al., Nature, 400(6746):761-766 (1999)]. Moreover, Korschen et al. demonstrated that GARP inhibits both soluble and membrane bound phosphodiesterase.

Cyclic nucleotides are involved in regulating the activity of airway smooth muscle and many other cells in the airways, including pro-inflammatory, immunocompetent cells such as macrophages, eosinophils, mast cells and lymphocytes. Cyclic nucleotides are inactivated by the action of cyclic nucleotide phosphodiesterase enzymes (PDE). Inhibition of cGMP PDE results in elevation of cGMP levels; elevated cGMP levels are associated with beneficial anti-platelet, anti-neutrophil, anti-vasospastic and vasodilatory activity.

Thus, the subject invention provides a polypeptide having the sequence of SEQ ID No: 306 or a GRP polypeptide encoded by the human cDNA of clone 187-38-0-0-110. In a preferred embodiment, GRP is encoded by the sequence of SEQ ID No: 65 or the human cDNA of clone 187-38-0-0-110, however, all polynucleotides encoding the polypeptides of the invention are included. As used herein, "the GRP protein" includes the full length protein of SEQ ID NO: 306 as well as biologically active fragments of the GRP protein. Also encompassed by the phrase "the GRP protein" are variants of the protein of SEQ ID NO: 306 and biologically active fragments of said variant proteins.

"Biologically active fragments" are defined as those peptide or polypeptide fragments of GRP which have at least one of the biological functions of the full length protein (*e.g.*, the ability to inhibit the activity of PDE or serve as an affinity substrate for PDEs).

The invention also provides variants of the protein of the GRP protein encoded by SEQ ID NO: 306. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence of GRP. Variants according to the subject invention also have at least one functional or structural characteristic of GRP, such as the ability to inhibit the activity of PDE or serve as an affinity substrate for PDEs. The invention also provides biologically active fragments of the variant proteins. Unless otherwise indicated, the methods disclosed herein can be practiced utilizing GRP or variants thereof. Likewise, the methods of the subject invention can be practiced using biologically active fragments of GRP, or biologically active fragments of GRP variants.

Assays related to the inhibitory effect of the protein of the invention can be carried out using techniques described in U.S. Patent No. 6,130,333, hereby incorporated by reference in its entirety; or by any other technique known to those skilled in the art.

One aspect of the subject invention provides compositions and methods of using the nucleotide sequence of SEQ ID NO: 65, or its complement, in molecular biology techniques. These techniques include, but are not limited to: the use of segments of GRP2 as oligomers for PCR; expression of the GRP2 and the production of recombinant proteins; in generation of antisense RNA and DNA, their chemical analogs and the like; the use of GRP2 segments as hybridization probes and in chromosome gene mapping.

For example, nucleotide sequence of SEQ ID No: 65, or its complement, can be used to generate hybridization probes for mapping the naturally occurring genomic sequence. The sequence can be mapped to a particular chromosome or to a specific region of the chromosome using well known techniques. These include in situ hybridization to chromosomal spreads, flow-sorted chromosomal preparations, or artificial chromosome constructions such as yeast artificial chromosomes, bacterial artificial chromosomes, bacterial P1 constructions, or single chromosome cDNA libraries as reviewed in Price (Price CM – Blood Rev. – 1993, 7(2):127-34) and Trask B (Trask BJ – Trends Genet. – 1991, 7(5):149-54).

In situ hybridization of chromosomal preparations and physical mapping techniques, such as linkage analysis using established chromosomal markers, are invaluable in extending genetic maps; genetic maps provide valuable information to investigators searching for disease-causing genes using positional cloning or other gene discovery techniques. The nucleotide sequence of the present invention can also be used to detect differences in the chromosomal location due to translocation, inversion, etc. among normal, carrier or affected individuals.

Another embodiment of the subject invention provides pharmaceutical compositions comprising the GRP protein and pharmaceutically acceptable carriers. These pharmaceutical compositions can be used in prophylaxis and/or treatment of a variety of conditions where inhibition of phosphodiesterase is considered to be beneficial. The biochemical, physiological, and clinical effects of phosphodiesterases inhibitors suggest their utility in a variety of disease states in which modulation of smooth muscle, renal, hemostatic, inflammatory, and/or endocrine function is desirable. Therefore, the GRP protein can be used for the treatment or prophylaxis of a number of disorders and conditions including, but not limited to, stable, unstable, and variant (Prinzmetal) angina; hypertension; pulmonary hypertension; congestive heart failure; acute respiratory distress syndrome; acute and chronic renal failure; atherosclerosis; conditions of reduced blood vessel patency (e.g., postpercutaneous transluminal coronary or carotid angioplasty, or post-bypass surgery graft stenosis); peripheral vascular disease; vascular disorders, such as Raynaud's disease, thrombocythemia, intermittent claudication; immune diseases, multiple sclerosis; cancers inflammatory diseases, graft versus host disease, Alzheimer's disease, memory deficits, , stroke,

bronchitis, chronic asthma, acute lung injury, chronic obstructive pulmonary disease, allergic asthma, allergic rhinitis; glaucoma; osteoporosis; preterm labor; benign prostatic hypertrophy; male and female erectile dysfunction; and diseases characterized by disorders of gut motility (e.g., irritable bowel syndrome).

The GRP protein of the invention can also provide beneficial anti-platelet, anti-neutrophil, anti-vasospastic, vasodilatory, natriuretic, and diuretic activities when administered in therapeutically effective amounts. The GRP protein can also potentiate the effects of endothelium-derived relaxing factor (EDRF), gastric NO administration, nitrovasodilators, atrial natriuretic factor (ANF), brain natriuretic peptide (BNP), C-type natriuretic peptide (CNP), and endothelium-dependent relaxing agents such as bradykinin and acetylcholine, when administered to an individual.

Another embodiment of the subject provides methods of treating male erectile dysfunction comprising the administration of therapeutically effective amounts of the GRP protein using appropriate methods known to the skilled artisan.

Another embodiment of the subject invention provides industrially significant methods of recovering PDE comprising contacting solutions containing PDE with immobilized GRP protein. In this aspect of the invention, the GRP protein is immobilized onto a solid support and allowed to specifically bind to PDE contained in a solution or sample. PDE can then be eluted from the immobilized GRP protein according to methods known to the skilled artisan (see, for example, Korschen et al., supra). PDE is a commercially valuable commodity sold by various vendors.

Protein of SEQ ID NO: 302 (internal designation 187-2-2-0-A3-CS)

The protein of SEQ ID NO: 302 encoded by the cDNA of SEQ ID No: 61 is related to a neuronally expressed protein (neuritin, Genseq accession number W37859) known to have a role in neuritigenesis and axonal and dendritic growth.

The 164 amino acid protein of SEQ ID NO: 302 is 24% identical to neuritin over the complete sequence. Specifically, SEQ ID NO: 302 displays two blocks of strong homology to neuritin (amino acids 41-60 of SEQ ID NO: 302 display 55% identity and 95% similarity to amino acids 30-49 of neuritin, and amino acids 66-117 of SEQ ID NO: 302 display 32% identity and 57% similarity to amino acids 62-113 of neuritin). The C-terminal portion of neuritin (aa 116-142 of neuritin) is highly hydrophobic and contains a cleavage site found in GPI-anchored proteins. The protein of SEQ ID NO: 302 also has a hydrophobic C-terminus (21 out of the last 30 amino acids are hydrophobic) and conforms to the GPI anchor consensus sequence.

Neuritin, also known as candidate plasticity-related gene number 15 (cpg-15), was independently identified by two groups from differential cDNA libraries generated from kainic acid-treated hippocampal cells (Nedivi et al., Nature. 363:718-22 (1993); Naeve et al., Proc. Natl. Acad. Sci. USA. 94:2648-53 (1997)). Neuritin is a secreted protein that contains a potential GPI

anchoring domain believed to anchor the protein to the membranes of target cells. Neuritin is expressed strongly in the brain, and in particular, in systems with pronounced developmental plasticity, including the pyramidal neurons of the cornus ammons and the granule cells of the hippocampus dentate gyrus. Strong expression is also observed in layers of tenia tecta projecting to the olfactory bulb, the major target of the retinal ganglion cells, and the optical nerve layer of the superior colliculus (optic tectum); and localized expression is observed in the thalamic nuclei and the cerebral cortex (Nedivi et al., Pro. Natl. Acad. Sci. USA. 93:2048-53 (1996); Naeve et al., Proc. Natl. Acad. Sci. USA. 94:2648-53 (1997); Nedivi et al., Science. 281:1863-66 (1998)). mRNA is expressed throughout development and persists into adulthood. In addition, neuritin expression is upregulated in adults by brain derived neurotrophic factor (BDNF). Neuritin mRNA is also detected in the lung and the liver, although at lower levels than that observed in the CNS (Naeve et al., Proc. Natl. Acad. Sci. USA. 94:2648-53 (1997)).

Functional studies on the neuritin protein have revealed a role in neuronal growth. In one such study, rat cortical and hippocampal neurons were treated with recombinant forms of neuritin. Neurons treated with neuritin showed extensive neuritogenesis over control cultures. Specifically, neurons showed well-differentiated cell bodies with well-defined extensions after treatment with neuritin (Naeve et al., Proc. Natl. Acad. Sci. USA. 94:2648-53 (1997)). Other studies using frog optic tectum showed that transfection of tectum cells with neuritin cDNA can increase the growth rate of tectal cell dendrites (Nedivi et al., Science. 281:1863-66 (1998)). Studies have also shown that neuritin can modify the growth of retinotectal axons by increasing the elaboration of presynaptic axons and can promote the maturation of retinal tectal synapses (Cantallops et al., Nature Neuroscience. 3:1004-1011 (2000)). Together, these results indicate that neuritin promotes the growth of pre- and post-synaptic neurons and contributes to the formation and stabilization of mature synapses.

The subject invention provides the polypeptide of SEQ ID NO: 302 and polynucleotide sequences encoding the amino acid sequence of SEQ ID NO: 302. In one embodiment, the polypeptides of SEQ ID NO: 302, including fragments, variants, etc. are replaced by the corresponding polypeptide encoded by the human cDNA of clone 187-2-2-0-A3-CS. Also included in the invention are biologically active fragments of the protein of SEQ ID NO: 302 and polynucleotide sequences encoding these biologically active fragments. In another embodiment, biologically active fragments comprise amino acid positions 41-60, 66-117, 41-117, and 41-164. In another embodiment, these fragments may be joined together by chemical linkers or by recombinantly inserted amino acid linker segments according to methods known in the art. "Biologically active fragments" are defined as those peptide or polypeptide fragments of SEQ ID NO: 302 which have at least one of the biological functions of the full length protein (e.g., the ability to stimulate neuritigenesis and axonal and dendritic growth).

The invention also provides variants of SEQ ID NO: 302. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence of SEQ ID NO: 302. Variants according to the subject invention also have at least one functional or structural characteristic of SEQ ID NO: 302, such as the biological functions described above. The invention also provides biologically active fragments of the variant proteins. Unless otherwise indicated, the methods disclosed herein can be practiced utilizing the polypeptide of SEQ ID NO: 302 or variants thereof. Likewise, the methods of the subject invention can be practiced using biological fragments of the protein of SEQ ID NO: 302 or variants of said biologically active fragments.

Because of the redundancy of the genetic code, a variety of different DNA sequences can encode SEQ ID NO: 302. It is well within the skill of a person trained in the art to create these alternative DNA sequences, which encode proteins having the same, or essentially the same, amino acid sequence. These variant DNA sequences are, thus, within the scope of the subject invention. As used herein, reference to "essentially the same sequence" refers to sequences that have amino acid substitutions, deletions, additions, or insertions that do not materially affect biological activity. Fragments retaining one or more characteristic biological activity of SEQ ID NO: 302 are also included in this definition.

"Recombinant nucleotide variants" are alternate polynucleotides which encode a particular protein. They can be synthesized, for example, by making use of the "redundancy" in the genetic code. Various codon substitutions, such as the silent changes which produce specific restriction sites or codon usage-specific mutations, can be introduced to optimize cloning into a plasmid or viral vector or expression in a particular prokaryotic or eukaryotic host system, respectively.

The protein of SEQ ID NO: 302, and variants thereof, can be used to produce antibodies according to methods well known in the art. The antibodies can be monoclonal or polyclonal. Antibodies can also be synthesized against immunogenic fragments of SEQ ID NO: 302, as well as variants thereof, according to known methods. The subject invention also provides antibodies which specifically bind to biologically active fragments of SEQ ID NO: 302 or biologically active fragments of SEQ ID NO: 302 variants.

The protein of SEQ ID NO: 302 can be utilized to treat diseases and disorders of the central or peripheral nervous system which arise from alterations in the pattern of expression of the protein of SEQ ID NO: 302. In this aspect of the subject invention, compositions comprising the protein of SEQ ID NO: 302 and a pharmaceutical carrier are administered to an individual in need thereof. Alternatively, in cases where the protein of SEQ ID NO: 302 is overexpressed, reductions in SEQ ID NO: 302 levels may be accomplished by a variety of methods known to those of skill in the art. These methods include the introduction of neutralizing antibodies or the use of antisense polynucleotides derived from the protein of SEQ ID NO: 302 or clone 187-2-2-0-A3-CS.

The subject invention also provides materials and methods for the treatment of neurological disorders comprising contacting neuronal cells with compositions comprising the protein of SEQ ID NO: 302 and pharmaceutically acceptable carriers. Thus, this aspect of the invention provides methods of treating patients suffering from a variety of neurological disorders, conditions, and/or diseases of the central, autonomic, or peripheral nervous system. These include neurological damage arising from congenital disease, trauma, surgery, stroke, ischemia, infection, metabolic disease, nutritional deficiency, malignancy, and/or exposure to toxic agents. Additional examples of such disorders include, but are not limited to, epilepsy, cerebral neutralisms, Alzheimer's disease, Pick's disease, Huntington's disease, dementia, Parkinson's disease and other extrapyramidal disorders, amyotrophic lateral sclerosis and other motor neuron disorders, progressive neural muscular atrophy, retinitis pigmentosa, hereditary ataxias, multiple sclerosis and other demyelinating diseases, bacterial and viral meningitis, brain abscess, subdural empyema, epidural abscess, auppurative intracranial thrombophlebitis, myelitis and radiculitis, viral central nervous system disease, prion diseases, Creutzfeldt-Jakob disease, Gerstmann-Staussler-Scheinker syndrome, fatal familial insomnia, diabetes induced peripheral neuropathy or neuropathy induced by other metabolic disorders or nutritional deficiencies, neurofibromatosis, tuberous sclerosis, cerebelloretinal hemangioblastomatosis, encephalotrigeminal syndrome, mental retardation and other developmental disorders of the central nervous system, cerebral palsy, neuroskeletal disorders, autonomic nervous system disorders, cranial nerve disorders, spinal cord diseases, muscular dystrophy and other neuromuscular disorders, dermatomyositis and polymyostis, inherited, metabolic, endocrine, and toxic myopathies, myasthenia gravis, periodic paralysis, mental disorders including mood, anxiety, and schizophrenic disorders, seasonal affective disorder, akathesia, amnesia, and/or other dystrophies or degenerative disorders of the visual, sensory, olfactory, auditory, motor, or memory systems. Methods of introducing therapeutic compounds into cells are known to those skilled in the art. Non-limiting examples include the use of targeted liposomes, fusogenic liposomes, or other carriers suitable for the introduction of a therapeutic compound into a target cell.

The subject invention also provides materials and methods for the treatment of neurological disorders comprising contacting neuronal cells with compositions comprising polynucleotides encoding the protein of SEQ ID NO: 302 and pharmaceutically acceptable carriers. In one embodiment, the polynucleotide is clone 187-2-2-0-A3-CS. Methods of introducing polynucleotides into cells and directing expression of the polynucleotide are known to those skilled in the art.

Antibodies raised against the protein of SEQ ID NO: 302 may be used in a variety of immunoassays known to those skilled in the art. In this aspect of the invention, immunoassay screening for abnormal levels of the protein of SEQ ID NO: 302 can be used as screens or diagnostic/prognostic indicators of neurodegenerative disease.

Antibodies raised against the protein of SEQ ID NO: 302, fragments, and/or derivatives thereof may also be used for detection and identification of growing and differentiating neurons including, but not limited to, the pyramidal neurons of the cornus ammons, the granule cells of the hippocampus dentate gyrus, neurons in layers of tenia tecta projecting to the olfactory bulb, the optical nerve layers of the superior colliculus (optic tectum), and neurons of the thalamic nuclei and the cerebral cortex.

Protein of SEQ ID NO:301 (187-12-4-0-A8-CS)

The protein of SEQ ID No:301, encoded by the cDNA of SEQ ID NO:60, is homologous to the Eukaryotic cell growth inhibiting factor (GENESEQP: R95950) described in patent WO9617933. The protein of the invention is highly expressed in the brain, fetal brain, fetal liver and the prostate.

It is believed that the protein of the invention is a cell growth inhibiting factor. Preferred polypeptides of the invention are those that comprise amino acids 221 to 287. Other preferred polypeptides of the invention are any fragment of SEQ ID NO:301 having any of the biological activities described herein. In the present invention, a cell inhibiting factor is defined as a peptide or protein that decreases, suppresses or terminates (reversibly or irreversibly) the growth of at least one type of cell such as, but not limited to, bacteria, yeast, vertebrate cells, mammalian cells and human cells, under ordinary culturing conditions known to those skilled in the art. Assay of the inhibiting activity of the invention can be carried out, for example, by evaluating the decrease in DNA synthesis as described in Patent WO 96/17933, or by measuring the number or density of cells using any standard method. For example, fibroblasts are transfected with a vector containing the DNA sequence coding for the protein of the invention or part thereof. Cells are then cultured in a standard medium, exposed to tritiated thymidine, and further cultured. The cultures are then fixed and stained with X-Gal, the blue stained galactosidase-expressing cells are counted under a microscope, and the ratio of cells showing dark particles in their nuclei due to tritiated thymidine uptake is determined. DNA synthesis inhibitory rates are calculated with the labeling index taking for reference (i.e. 0% inhibition) a culture of cells tranfected with a "blank" vector (i.e. not modified to contain the DNA coding for the protein of the invention or part thereof).

Aging at the cell level is associated with individual aging. The maximum possible number of divisions (division life span) of cultured cells is inversely proportional to individual age. Even if an aged cell is fused with a young or immortalized cell, DNA synthesis does not occur again in aged cells; on the contrary, DNA synthesis in the young and immortalized cell is suppressed (Stein GH et al. – Proc Nat Acad Sci. – 1981, 78:p3025). This demonstrates that certain factors controlling cellular senescence are dominant, and that aged cells not only lack substances essential for their growth but also have substances that actively suppress DNA synthesis. Moreover, microinjections of mRNA, prepared from an aged cell, are known to inhibit DNA synthesis (Lumpkin CK et al. –

Science – 1986, 232:p393). Therefore as cells age, there are some genes that are newly expressed or whose expression is increased. Such genes play an important role, directly or indirectly, in cell aging.

Pereira-Smith et al. tested the complementation of a large number of immortalized human cells in fused pairs and demonstrated the presence of 4 groups of human aging genes (Pereira-Smith OM et al. – Proc Nat Acad Sci. – 1988, 85:p6042). Clarifying the nature of aging-associated genes is not only important in understanding aging, both at the cellular and individual levels, but is also significant in that the use of these genes or gene products would enable the diagnosis of various aging-associated diseases and diseases caused by cellular senescence, the development of prophylactic/therapeutic drugs for such diseases, and their application as prophylactic/therapeutic drugs for various diseases involving uncontrolled cell growth such as, but not limited to, cancers.

In one embodiment of the present invention, the polypeptides and polynucleotides of the invention are used to specifically label cells of the brain, fetal brain, fetal liver and the prostate, as the protein is strongly expressed in these tissues. The ability to specifically detect these tissues, and cells derived from these tissues, has a number of uses, including for the determination of the history of tumor cells and for histological analyses.

An embodiment of the present invention invention relates to methods and compositions of using the protein of SEQ ID NO:301 or the cDNA of SEQ ID NO:60 or any part thereof, to inhibit cell proliferation in vitro. For example, by including the invention in a "cocktail" with other proteins (such as proteases) it could be used as a decontaminant, i.e. to prevent the growth of any cells to maintain a sterile environment. Preferred applications of this embodiment include decontamination of samples (such as cell culture media) and instruments (such as surgical instruments), where the invention would be used as a bacteriostatic/mycostatic agent. Another example pertains to the use of the protein of the invention as a reagent for terminating the cell cycle of cultured cells at a given time point, e.g., as a reagent for synchronizing cell division, avoiding the need to isolate specific cells (e.g. at the desired cell cycle phase) in cultures using techniques such as flow cytometry. Synchronization of the cell cycle could, for example, be achieved, e.g., by transfecting cells with an appropriate vector containing the DNA coding for the protein of the invention or part thereof, where expression of the protein results in growth inhibition. Then, after a certain time, an inhibitor of the protein could be administered in order to enable the cells to resume growth (e.g. all at the S phase, when DNA is synthesized). Further, the ability to synchronize the cell cycle in an in vitro experimental system would provide improved assay precision or would facilitate any laboratory procedure or experiment involving a particular cell cycle stage. Use of the invention for in vitro inhibition of cell proliferation is not limited to the above examples; the invention is potentially useful in any in vitro application that requires the inhibition of cellular proliferation.

Another embodiment of the invention pertains to the introduction of SEQ ID NO:301 or SEQ ID NO:60, or any part thereof, into a target tissue (such as skin or vascular endothelium) to establish an in vitro aged cell line of the target tissue. Such a cell line is useful as a screening system for clarifying the mechanisms of aging and/or cellular senescence, but also for seeking prophylactic and therapeutic drugs for aging-associated diseases and or diseases caused by cellular senescence, as it provides an in vitro model of aged cells (in vitro cell cultures provide the advantage of being easily produced and are not as expensive as animal models). Thus these cells are potentially useful in drug candidate screening applications; a preferred application involves its use in high-throughput screening, e.g. to identified "lead compounds."

A preferred embodiment of the invention relates to the use of the cDNA of SEQ ID NO:60 or part thereof, as a probe for examining individual aging at the gene expression level. Specifically, SEQ ID NO:60, or part thereof, can be used as a diagnostic reagent for various aging-associated diseases such as, but not limited to: arteriosclerosis, osteo-arthritis, dementia (including Alzheimer's disease) and Parkinson's disease.

In a related embodiment, the cDNA of SEQ ID NO:60 or part thereof, could be used to synthesize antisense oligonucleotides by methods well known to those skilled in the art. Antisense oligonucleotides can be used to inhibit the synthesis of the protein of SEQ ID NO:301, thereby preventing cell and tissue aging and/or promoting the rejuvenation of aged cells and tissues. These antisense oligonucleotides can also be used for in vivo or ex vivo treatment and prophylaxis of diseases caused by cellular senescence or aging-associated diseases such as, but not limited to: arteriosclerosis, osteo-arthritis, dementia (including Alzheimer's disease) and Parkinson's disease.

In a most preferred embodiment, SEQ ID NO:301, SEQ ID NO:60, or any part thereof, can be used as a pharmaceutical drug to treat pathologies such as, but not limited to cancers, inflammation, or infections. For example, when used as an antibacterial, antiviral and/or antifungal agent, inhibition of microbial proliferation could be achieved by either directly inhibiting microorganism growth (in the case of fungal and bacterial infections) or DNA synthesis of infected cells in the case of viral infections. The DNA of SEQ ID NO:60 or part thereof, can also be used to develop gene therapy products in the in vivo or ex vivo treatment of diseases and conditions such as cancers and inflammation. SEQ ID NO:301, SEQ ID NO:60, or any part thereof, may also be used: 1) as a probe for the diagnosis, 2) as a prophylaxis or 3) as a treatment of aging-associated diseases such as, but not limited to: arteriosclerosis, osteo-arthritis, dementia (including Alzheimer's disease) and Parkinson's disease. In a related embodiment, the protein of SEQ ID NO:301 or the cDNA of SEQ ID NO:60, or any part thereof, can be used in the development of drugs that will be used in the prophylaxis or treatment of the diseases stated above (e.g. diseases caused by cellular senescence, aging-associated diseases and diseases caused by cellular proliferation).

For the treatment or prevention of diseases and conditions associated with undesired proliferation, such as cancer, inflammation, or infection, the expression or activity of the present

protein can be increased using any of a number of methods. For example, polynucleotides encoding the protein can be introduced into the undesired cells, wherein the protein is then expressed and inhibits the further growth of the cells. In one such embodiment, the polynucleotides can be incorporated into liposomes comprising on their surface a specific molecule that directs the targeting of the liposome to a specific cell type (e.g. a tumor-specific antibody). Alternatively, the protein of SEQ ID NO:301 can itself be administered to the cells, e.g. as a fusion protein also comprising a specific targeting polypeptide moiety. Further, a compound that enhances the expression or activity of the protein can be administered to cells, preferably in a way that specifically targets the compound to undesired cells, e.g. chemically linked to a heterologous specific targeting molecule.

Protein of SEQ ID NO : 412 (internal designation 187-5-3-0-C7-CS)

The protein of SEQ ID NO : 412 encoded by the cDNA of SEQ ID NO : 171 is homologous to the human CDK4-binding protein p34 $^{SEI-1}$ (sptrembl accession number Q9UHV2). p34^{SEI-1} is a new CDK4 regulator that prevents p16INK4a from inhibiting the formation of cyclinD1-CDK4 complexes. p34^{SEI-} seems to act as a growth factor sensor and may facilitate the formation and activation of cyclin D-CDK complexes in the face of inhibitory levels of INK4 proteins (Sugimoto et al., Genes Dev. 13:3027-3033 (1999)).

Progression through the cell cycle is a complex process that is regulated at many levels by several proteins. The activity of cyclin dependent kinases (CDK4 and CDK6) is regulated by the association of cyclin partner that acts as a positive effector and by two families of cdk inhibitors proteins (KIP) and the inhibitors of cdk4 (INK4) such as p16INK4a, which act as negative effectors (Sandhu et al., Cancer Detect. Prev.24:107-118 (2000)).

Cancer is a disease characterized by loss of cellular growth control, the molecular machinery of the cell cycle is involved in tumorigenesis. Many human tumors have been shown to have abnormality in this pathway resulting in either the functional inactivation of p16INK4a or the excessive activity of CDK4 (Palmero at al., Cancer Surv.27:351-357 (1996)).

It is believed that the protein SEQ ID No: 412 plays a role in the cell cycle regulation via the binding to a cyclin dependent kinase. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 412 having any of the biological activity described herein. The binding activity of the protein of the invention or part thereof to a cyclin dependent kinase, as well as its role in cell cycle, may be assayed using any of the assays known to those skilled in the art including those described in Sugimoto et al., supra.

An embodiment of the present invention relates to methods of using the protein of the invention or part thereof to identify and/or quantify cyclin dependent kinases, preferably CDK4, in a biological sample, and thus used in assays and diagnostic kits for the quantification of such CDKs

in bodily fluids, in tissue samples, and in mammalian cell cultures. The binding activity of the protein of the invention or part thereof may be assessed using the assay described in Sugimoto et al., supra or any other method familiar to those skilled in the art. Preferably, a defined quantity of the protein of the invention or part thereof is added to the sample under conditions allowing the formation of a complex between the protein of the invention or part thereof and the cyclin dependent kinase to be identified and/or quantified. Then, the presence of the complex and/or or the free protein of the invention or part thereof is assayed and eventually compared to a control using any of the techniques known by those skilled in the art.

In another embodiment, the invention relates to compositions and methods using the protein of the invention or part thereof to stimulate cell proliferation both in vitro and in vivo. For example, soluble forms of the protein of the invention or part thereof may be added to cell culture medium in an amount effective to stimulate cell proliferation.

The invention further relates to methods and compositions using the protein of the invention or part thereof to diagnose, prevent and/or treat several disorders associated with cell proliferation including but are not limited to, adenocarcinoma, sarcoma, lymphoma, leukemia, melanoma, myeloma, teratocarcinoma, cancers of the adrenal gland, bladder, bone, brain, breast, gastrointestinal tract, heart, kidney, liver, lung, ovary, pancreas, paraganglia, parathyroid, prostate, salivary gland, skin, spleen, testis, thyroid, uterus, and neurodegenerative disorders such as Alzheimer's disease (McShea et al., Am.J.Pathol.150(6):1933-1939 (1997)). For diagnostic purposes, quantification of the protein of the invention could be investigated, using Northern blotting, RT-PCR, immunoblotting and any of protocols known in the art, in biological samples and compared to the expression in control biological samples. Thus a diagnosis assay may be used, to determine altered expression of the protein of the invention, to correlate with diseases states and to evaluate the prognostic significance in diseases. For prevention and/or treatment purposes, inhibition of the endogenous expression of the protein of the invention using any of the antisense or triple helix methods described herein may be used. Alternatively, inhibitors for the protein's activity may be developed and use to inhibit and/or to reduce the protein's activity using any methods known to those skilled in the art. Antibodies which specifically bind to the protein of the invention may be generated using methods that are well known in the art and used as an antagonist.

Protein of SEQ ID NO : 299 (internal designation 184-1-4-0-C11-CS)

The protein of SEQ ID NO : 299 encoded by the cDNA of SEQ ID NO: 58 and found in fetal liver and liver, is orthologous to the BolA protein. The BolA family comprises the morpho-protein BolA from *E. coli* and its various homologs. The expression of BolA is growth rate regulated and is induced during the transition into the stationary phase. BolA is also induced by stress during early stages of growth and can have a general role in stress response. It has also been

suggested that BolA can induce the transcription of penicillin binding protein 6 and 5 (EMBO J. 1;8 (2) :3923-31 (1989)).

E. coli cells become thinner and shorter after a period of starvation or stationary-phase conditions; this altered morphology is an adaptative response of *E. coli* to general forms of stress. The bolA gene seems to be involved in the switching between cell elongation and septation systems during the cell division cycle [J Bacteriol 170 :5169-5176 (1988)]. The regulation of bolA has been linked to the presence of gearbox promoter from which RNA is transcribed [Mol Microbiol 5 :2085-2091 (1991)].

Expression of bolA is governed by two promoters. P2 is located further upstream from the structural gene, is under the control of σ^d and transcribes bolA constitutively. The promoter P1, proximal to the structural gene, is a gearbox promoter under the control of $\sigma\square^s$ from which bolA has been shown to be transcribed in an inverse growth rate-dependent fashion [J Bacteriol 173 :4474-4481 (1991)].

The alternate sigma factor $\sigma\square^s$ is encoded by the gene rpoS and has been described as a central regulator for the induction of a set of specific genes involved in adaptation to stationary phase. It has, nevertheless, been shown that $\sigma\square^s$ function is not confined to stationary phase. Significant increases in $\sigma\square^s$ cellular levels were seen during exponential growth in response to forms of stress; genes under its control code for important adaptive regulators for general stress conditions [FEMS Microbiol Lett 30 :419-430 (1997)].

The smaller morphology caused by stress-induced overexpression of bolA reduces the surface area exposed to the environment and decreases the cell's surface-to-volume ratio.

Identification of ortholog genes provides important information regarding functional and structural conservation within these orthologs throughout evolution. The concept of comparative gene identification has been previously used by many laboratories to search for orthologous genes once a particular gene of interest has been identified in another species [Genome Res 10 (5) : 703-13 (2000)].

The protein of invention contains a signal peptide corresponding to a short helix as predicted by software TopPred II [Clarosand and von Heijne, CABIOS applic. Notes, 10: 685-686 (1994)]. Thus, one aspect of this invention provides materials and methods for the delivery of recombinant proteins to liver cells. The signal peptide, encoded by an appropriate polynucleotide, can be linked to another protein/polypeptide (also encoded by an appropriate polynucleotide). The recombinant gene, containing the signal peptide sequence, is expressed and the desired protein is delivered via the signal peptide. Methods of producing such gene fusions, the expression of such gene products, and their use are well known to the skilled artisan.

In another embodiment, BolA, or biologically active fragments thereof, can be used to modulate the stress response to environmental changes such as cytotoxic agents, heat shock, irradiation, genotoxic stress or growth factors.

In another aspect of the invention, SEQ ID NO: 299 is incorporated into a prokaryotic expression vector and transfected into prokaryotic cells unable to adapt to environmental stress or cells containing a BolA defect. The expression vector can, optionally, contain a promoter system such as that described supra (e.g., P1, P2, σ^d, σ^s, etc.) which typically controls bolA expression. The components necessary for transcription can be provided in one or more expression vectors. Prokaryotic cells thus transformed can be useful in bioremediation systems where environmental stress is commonly encountered. Thus, preferred prokaryotes for the practice of this aspect of the invention lack bolA, or contain a bolA defect, and are known to be useful for bioremediation.

In another embodiment, the subject invention provides methods and compositions to selectively identify liver tissues. The protein encoded by SEQ ID NO: 299 can be used to synthesize specific polyclonal or monoclonal antibodies using any techniques known to those skilled in the art. These antibodies can be used to selectively identify liver tissue according to well-known histological immunoassays. The ability to immunologically identify tissue samples is industrially important for analysis of mismarked biopsy samples (e.g., laboratory errors) where the origin of the tissue sample is in question, or simply to verify that a tissue sample originated from liver. The antibodies can also be used to identify cancer metastases originating from the liver.

Further, antibodies provided by the subject invention can also be used to assay animal feeds for the presence of liver or liver by-products. As is known, many animal feeds contain animal protein. The use of animal feeds containing animal protein has been associated with disorders in both animals and humans (the most notorious of which is bovine spongiform encephalitis). This has resulted in the banning of animal protein in feeds provided to animals. However, to ensure compliance with such bans, animal feeds must be tested. Thus, the antibodies of the invention can be used to test animal feeds for the presence of liver according to methods known to those skilled in the art.

Proteins of SEQ ID NOs: 249 and 288 (internal designation 105-037-2-0-H11-CS and 174-7-4-0-H1-CS respectively).

The 403-amino-acid-long protein of SEQ ID NO: 249 encoded by the cDNA of SEQ ID NO: 8 is extensively homologous to the protein of SEQ ID NO: 288 encoded by the cDNA of SEQ ID NO: 47 with the exception of five amino acids in positions 192-194, and 298-299 which are not present in protein of SEQ ID NO: 288. It is likely that the two proteins are the result of an alternative splicing and display similar functions and utilities.

The 403-amino-acid-long protein of SEQ ID NO: 249, overexpressed in salivary glands, exhibits extensive homology to the mus musculus hypothetical protein (Genbank accession number AB030196). The amino acid residues of protein of SEQ ID: 249 show a high degree of identity to the Genbank sequence. However, the protein of Genbank sequence does not have the twenty amino acids (192 to 194, and 298-303, 353-354, 380-381, and 387-393) and also displays 35 different

amino acids from the SEQ ID NO: 249. In addition, four transmembrane domains are predicted for the protein of SEQ ID NO: 249 from positions 31 to 51, 75 to 95, 154 to 174, and from 236 to 256 as predicted by the software TopPred II (Claros and von Heijne, *CABIOS applic. Notes*, 10 : 685-686 (1994)).

When expressed in *E. Coli*, the matched sequence suppresses bacterial growth (Inoue et al, Biochem Biophys Res Commun 268:553-61 (2000)). It is therefore believed that the proteins of SEQ ID NO: 249 and 288 or a bacterial growth suppressing fragment thereof can be used to suppress bacterial growth by contacting bacteria (gram negative or gram positive) with the polypeptides of the invention. The growth inhibiting activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in Inoue et al, supra.

In accordance with one aspect of the invention, methods and compositions using the protein of the invention or a fragment thereof to suppress bacterial growth are provided. In a preferred embodiment, the protein of the invention is expressed in a bacteria, preferably *E. coli*, using recombinant DNA technology methods known to those skilled in the art. The expressed protein can then be used to inhibit bacterial growth. The effects of the expressed protein and analogs or antagonists thereof can be assessed using any methods or techniques known to those skilled in the art.

Further included in the invention are the polypeptides encoded by the human cDNA of clone 105-037-2-0-H11-CS-SD. The polypeptides of SEQ ID NO: 249 may be interchanged with the corresponding polypeptides encoded by the human cDNA of clone 105-037-2-0-H11-CS-SD. Further included in the invention are polynucleotides encoding said polypeptides. Preferred polynucleotides are those of SEQ ID NO: 8 and of the human cDNA of clone 105-037-2-0-H11-CS-SD.

Nucleotide sequences encoding the polypeptides of SEQ. ID. NOs: 249 and 288 can be used to generate probes for the detection of related genes. Vectors expressing the nucleotide sequence can be used can be used to express the polypeptide in target cells. Antisense nucleotides can be used to inhibit the expression of the polypeptide.

Thus, in another embodiment of the invention the protein or a fragment thereof can be used as a marker protein to selectively identify tissues, preferably salivary glands. For example, the protein of the invention or a fragment may be used to synthesize specific antibodies using any techniques known to those skilled in the art. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry. In another embodiment, polynucleotides encoding the protein of SEQ. ID. NO. 249 can be used for *in situ* hybridization.

The transcript coding for the protein gng3lg (Genebank accession number AF069954) is transcribed from a bidirectional promoter divergently with the transcript coding for the gamma 3 subunit protein called gng3, a novel human G binding protein gamma-3 (HGPG) (Genbank accession number AF069953) and this organization is conserved across species within the human genome (*Downes et al, Genomics, 53:220-230 (1998)).*

Several genes which are linked in common physiological functions share a common divergently bidirectionnel promoter like αB crystallin and α crystallin, collagen type IV A1 and A2, surf1-3, and surf 5 genes, dihydrofolate reductase and 2 mismatch repair1 *(Iwaki et al., Genomics, 45: 386-394 (1997), Burbelo et al., Proc. Natl. Acad. Sci. USA 85: 9679-9682 (1988), Kaytes et al., J. Biol. Chem, 263: 19274-19277 (1988), Poschl et al., EMBO J., 7:2687-2695 (1988), Soininen et al., J. Biol. Chem, 263: 17217-17220 (1988), Garson et al., Genomics, 30: 163-170 (1995), Williams et al., Mol. Cell. Biol., 6: 4558-4569 (1986), Fujii et al., J. Biol. Chem., 264: 10057-10064 (1989)).* The heterodimeric G proteins, a family of GTPases are present in all cells and control a variety of functions (metabolic, humoral, neural and developmental) by transducing hormonal, neurotransmitter and sensory signals into an array of cellular responses. Triggered by cell surface receptors, each G protein regulate the activity of a specific effector including adenylate cyclase, phospholipase C, and ion channels protein which initiate appropriate biochemical responses. In view of this, it is believed that the transcript coding for the proteins of SEQ ID NO: 249 shares common regulatory elements with gng3 gene and that the products of such genes which are protein of SEQ ID NO: 249 and gng3 are physiologically coupled in unknown ways. Thus, in an embodiment of the invention, the protein of SEQ ID NO: 249 or part thereof may be used to regulate signal transduction of hormonal, neurotransmitter, and sensory signals to provide an array of cellular responses.

In yet another embodiment of the invention, the polypeptides of the present invention and the related polynucleotides may be used to treat several types of disorders including, but not limited to, cancer, neurodegenerative diseases, cardiovascular disorders, hypertension, renal injury and repair, septic shock.

In one embodiment, the protein of SEQ. ID. NO: 249 or a fragment, derivative or analog thereof may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of the protein.

In a further embodiment, a vector capable of expressing the protein of SEQ. ID. NO: 249 may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of the protein. Naked nucleotides encoding the protein of SEQ. ID. NO 249 may also be used.

In yet another embodiment, a pharmaceutical composition comprising a substantially purified the protein of SEQ ID NO: 249 may be administered to a subject to treat or prevent a disorder associated with decreased expression of the protein.

In still another embodiment, the polypeptide of SEQ ID NO: 249 can be used to develop and screen antagonists. For example, purified polypeptide can be used to develop antibodies or to screen libraries of pharmaceutical agents to identify those that inhibit the physiological functions of the protein.

Thus, in a further embodiment, an antagonist of the protein of SEQ ID NO: 249 can be administered to a subject to prevent or treat a disease associated with increased expression or activity of the protein. Similarly, in another embodiment a vector expressing the complement of a polynucleotide sequence encoding the protein of SEQ ID NO: 249 may be administered to decrease the expression of the protein.

The protein of SEQ ID NO: 249 displays a leucine zipper pattern situated near its NH2 terminal part (position 20 to 41). Thus, it is believed that the protein of SEQ ID NO: 249 is able to dimerize either with itself (homo-dimerisation) or with an heterologous protein (hetero-dimerisation) of interest, through the mediation of its leucine zipper domain. Preferred polypeptides of the invention are polypeptides comprising leucine zipper domains fragments and fragments having any of the biological activities described herein. The multimerization activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including circular dichroism spectrum and thermal melting analyses as described in US patent 5,942,433. The utilities of proteins containing leucine zipper domains, such as the protein of SEQ ID No: 249, are described elsewhere in the application.

Protein of SEQ ID NO: 259 (internal designation 114-016-1-0-H8-CS)

The protein of SEQ ID NO: 259, herein referred to as HOPP, is encoded by clone 114-016-1-0-H8-CS (SEQ ID NO: 18). This protein is homologous to a protein of *Arabidopsis thaliana* (ASY1) and *Saccharomyces cerevisiae*, (HOP1) (Caryl A.P. et al. Chromosoma, 109, 62-71; Hollingsworth N.M. et al. Cell, 61, 73-84).

In addition, the 394-amino-acid protein of SEQ ID NO: 259 displays a pfam HORMA domain from position 22 to 230. The HORMA domain is a common structural element in mitotic checkpoints, chromosome synapsis and DNA repair. For example, the HORMA domain was found in: (1) MAD2, a key component of the mitotic-spindle-assembly checkpoint (reviewed in Straiht AF. Current Biology 1997, 7:613-616); (2) HOP1, a conserved protein that is involved in meiotic-synaptonemal-complex assembly (Hollingsworth N.M. et al. Cell, 61, 73-84); and (3) in Rev7p, a subunit of the yeast DNA polymerase "epsilon" that is involved in translation, template independent DNA synthesis (Aravind L. and Koonin E.V., Trends Biochem Sci. 1998 Aug;23(8):284-6).

The pairing of homologous chromosomes during meiotic prophase culminates in the formation of the synaptonemal complex (SC), which is a ribbon-like, proteinaceous structure that holds homologous chromosomes in close apposition along their entire length. The synaptonemal complex (SC) is a prominent and evolutionally well conserved structure which is strictly meiotic.

Evidence from mutant phenotypes supports the hypothesis that recombination and SC formation are mutually interdependent processes. First, although not required for homology recognition, the SC could promote interhomolog interactions in situations where the normal processes have failed (*e.g.*, interlocking, heterologous pairing, etc.). Second, polymerization of the SC components might permit the recombination process to progress by modulating the number and localization of reciprocal versus exchanges (i.e. interference). Third, the SC may play an important role in meiotic chromosome structure and especially inter-sister interactions.

Synapsis of homologous chromosomes is a key event in meiosis as it is essential for normal chromosome segregation and is implicated in the regulation of crossover frequency (for review see Zickler D., J Soc Biol 1999;193(1):17-22). Mutants in HOP1 and ASY1, both proteins having significant homology to the protein of SEQ ID NO: 259, display decreased levels of meiotic crossover and intragenic recombination between markers on homologous chromosomes (Hollingsworth N.M., Byers B., Genetics 1989 Mar;121(3):445-62 ; Caryl AP et al. Chromosoma 2000;109(1-2):62-71).

Thus, the invention relates to methods and compositions using the protein encoded by clone 114-016-1-0-H8-CS or polynucleotide of SEQ ID NO: 18, or biologically active fragments thereof, to restore normal chromosome segregation in cells by administration of compositions comprising HOPP polypeptide, or polynucleotide encoding a HOPP polypeptide, encoded by clone 114-016-1-0-H8-CS, or polynucleotide in therapeutically effective amounts. The loss of normal chromosome segregation in normal cells leads to aberrant chromosome segregation events, a hallmark of tumor progression. HOPP proteins, encoded by clone 114-016-1-0-H8-CS, can be targeted to the nucleus by nuclear targeting sequences according to well-known methods. Nuclear targeting sequences (or NLS) can be chemically or recombinantly attached to HOPP. Alternatively, the HOPP gene can be used in known gene therapy protocols to restore normal chromosome function.

Infertility due to gametogenic failure is frequently associated with structural autosomal abnormalities. Recent meiotic studies, at pachytene stage, have shown a failure around the breakpoints, an association of the translocation figure with the sex chromosomes, and the frequent involvement of the acrocentric chromosomes. Two main models are proposed to explain the male sterilizing effect of rearrangements. The impairment of spermatogenesis could be the result of: 1) the XY-autosome interaction; or 2) the disruption around the breakpoints at the pachytene stage. These defects may contribute significantly to germ-cell atresia (for review see Luciani JM, Guichaoua MR Reprod Nutr Dev 1990; Suppl 1:95s-103s and Miklos GL. Cytogenet Cell Genet. 1974;13(6):558-77). Thus the subject invention also relates to methods and compositions of using the protein of SEQ ID NO: HOPP or clone 114-016-1-0-H8-CS, or biologically active fragments thereof, to reduce the incidence of infertility due to gametogenic failure. HOPP can be introduced into sperm as described in the preceding paragraphs. HOPP, optionally joined to a NLS sequence,

can also be introduced into sperm or eggs by other methods well known in the art (such as electroporation or microinjection).

The protein of SEQ ID NO: 259, encoded by clone 114-016-1-0-H8-CS, can also arrest cell division in human cells if the mitotic spindle apparatus is improperly attached to the chromosomes (Allshire R.C. Current Opinion in Genetics and Development 1997, 7:264-273). In the absence of functional protein of SEQ ID NO: 259, cells exposed to drugs which inhibit the formation of a mitotic spindle, such as benomyl, vinblastine, nocodozole, etc. would be expected to undergo rapid cell death due to massive chromosome loss. Human cells containing HOPP would be expected to survive such drug treatment because they are able to stop dividing prior to the chromosome loss event. Tumor cells that are hypersensitive to chemotherapeutic agents, which inhibit the formation of the mitotic spindle, may be sensitive to these drugs because they are defective in the checkpoint protein. Thus, screening assays for the presence or absence of the protein in a given tumor would provide an indication of the chemosensitivity of a particular tumor. The present invention therefore includes methods of determining prognostic benefit of treating a patent with a chemotherapeutic agent or determining which chemotherapeutic agent from a group of at least two would a patient more likely benefit from. Furthermore, the loss of checkpoint function in a normal cell may predispose that cell to aberrant chromosome segregation events, a hallmark of tumor progression. Thus the antibodies, polypeptides and polynucleotides of the present invention would be useful in diagnosing particular cancers.

Polyclonal antibodies can be produced by injecting a host animal such as rabbit, rat, goat, mouse or other animal with an immunogen of this invention. The sera is extracted from the host animal and is screened to obtain polyclonal antibodies which are specific to the immunogen. Methods of screening for polyclonal antibodies are well known to those of ordinary skill in the art such as those disclosed in Harlow & Lane, Antibodies: A Laboratory Manual, (Cold Spring Harbor Laboratories, Cold Spring Harbor, N.Y.: 1988) the contents of which are hereby incorporated by reference.

The monoclonal antibodies can be produced by immunizing, for example, mice with an immunogen according to the invention. Methods of producing monoclonal antibodies are well-known in the art and include those methods Kohler, B. and Milstein, C., Nature (1975) 256: 495-497. Hybridomas can be expanded, if desired, and supernatants can be assayed by conventional immunoassay procedures, for example radioimmunoassay. Positive clones can be further characterized. Hybridomas that produce the desired antibodies can be grown in vitro or in vivo using known procedures. The monoclonal antibodies can be isolated by conventional immunoglobulin purification procedures such as ammonium sulfate precipitation, gel electrophoresis, dialysis, affinity chromatography, and ultrafiltration.

Antibodies of the invention can be labeled with a detectable moiety. As noted above, a "detectable moiety" is well known to those of ordinary skill in the art and include, but are not

limited to, a fluorescent label, a radioactive atom, a paramagnetic ion, biotin, a chemiluminescent label or a label which can be detected through a secondary enzymatic or binding step.

The invention further provides a method of determining the susceptibility of a tumor sample to treatment with a mitotic spindle inhibitor which comprises steps of: a) contacting the tumor sample with an antibody, wherein the antibody is labeled with a detectable moiety and is capable of specifically binding to the protein of invention, or a fragment thereof, and b) assaying for the presence of an immunocomplex formed in step (a). The absence of the immunocomplex indicates that the tumor would be susceptible to treatment with a mitotic spindle inhibitor such as benomyl, vinblastine, and nocodozole.

The subject invention also provides a pharmaceutical composition comprising nucleic acid encoding the protein of SEQ ID NO: 259 and a carrier. In one aspect of this invention, the compositions are capable of passing through a cell membrane and provide for the expression of the protein of invention. As used herein, the term "carrier" includes pharmaceutically acceptable carriers and encompasses any of the standard pharmaceutically accepted carriers, such as phosphate buffered saline solution, water, emulsions such as an oil/water emulsion or a triglyceride emulsion, various types of wetting agents, tablets, coated tablets and capsules. carriers contain excipients such as starch, milk, sugar, certain types of clay, gelatin, stensic acid, talc, vegetable fats or oils, gums, glycols, or other known excipients. Flavor and color additives or other ingredients can also be included. In addition to the standard characteristics of the pharmaceutically acceptable carriers, the "suitable" carriers of the subject can also include those carriers which are able to penetrate the cell membrane. Therefore in one embodiment of the pharmaceutical composition the pharmaceutically acceptable carrier binds to a receptor on a cell capable of being taken up by the cell after binding to the structure.

This invention further provides a method of suppressing tumor formation in a subject which comprises administering a nucleic acid encoding the protein of invention in an amount effective to enhance expression of this protein.

Proteins of SEQ ID NOs: 311 and 312 (internal designation 188-28-4-0-B12-CS.corr and 188-28-4-0-B12-CS.fr respectively)

The 466-amino-acid-long protein of SEQ ID NO: 311, encoded by the human cDNA of clone 188-28-4-0-B12-CS or the cDNA nucleotide sequence of SEQ ID NO: 70, is related to proliferating-cell nucleolar antigen p120 (Genbank accession number M32110) encoded by nol1; and the yeast nucleolar protein Nop2p coded by nop2 (Genbank accession number U12141). SEQ ID NO: 311 (encoded by clone 188-28-4-0-B12-CS.corr) shows strong homology with three proteins described as homologs of p120 (Genbank accession number AK002229 and Genesequ accession numbers: Y86441, Y86442). In addition, the protein of SEQ ID NO: 311 is a polymorphic variant of SEQ ID No: 312 encoded by the cDNA of SEQ ID No: 71.

In addition, the protein of the invention exhibits the pfam NOL1/NOP2/sun family signature from positions 201 to 276. This motif is also found by emotif from positions 230 to 245. The NOL1/NOP2/sun family include p120 and Nop2p. These proteins are involved in nucleolar structure and activity as well as the regulation of cell cycle.

Freeman J.W. et al. (Cancer Res. 48: 1244-51, 1988) identified p120, a 120-kD nucleolar antigen associated with proliferating cells. This protein is a proliferation-associated antigen that is temporally regulated during the cell cycle and demonstrates a dramatic increase in expression at the G1-S boundary. This suggests that p120 can play a role in the regulation of the cell cycle and the increased nucleolar activity that is associated with cell proliferation (Fonagy A. et al. (1993) J. Cell. Physiol. 154:16-27).

The human p120 protein is also the most cancer specific of the identified proliferation-associated nucleolar proteins. Antigen p120 was detectable in a broad range of human malignant tumors but not in benign tumors or corresponding normal tissues. The antigen was not detectable in growth-arrested cells but was expressed early in G1 of the cell cycle.

Overexpression of human p120 leads to the transformation of NIH 3T3 cells. Expression of antisense p120 constructs causes the p120-transformed cells to revert to their original phenotype. Perlaky L. et al. [Cancer Res. 52: 428-36 (1992)] and Valdez et al. [Cancer Res. 52: 5681-87 (1992)] reported that the middle region of antisense p120 RNA inhibited proliferation of NIH 3T3 cells to approximately the same extent as the full-length antisense construct. The predicted mouse and human P120 proteins are 63% identical.

Another protein of the Nol1/Nop2/Sun family, Nop2p, coded by the gene NOP2, has a role in nucleolar function during the onset of growth and in the maintenance of nucleolar structure (de Beus et al. (1994) J. Cell Biol. 127:1799-813). The two proteins, p120 and Nop2p, are associated to ribosomal RNA in pre-ribosomal particles and can mediate the maturation process of the ribosome (Hong B. et al. (1997) Mol. Cell Biol. 17:378-88; Gustafson W.C. et al. (1998) Biochem. J. 331:387-93).

The subject invention provides the polypeptides encoded by the human cDNA of clone 188-28-4-0-B12-CS and polynucleotide sequences encoding the same amino acid sequences. Also included in the invention are biologically active fragments of the protein encoded by the human cDNA of clone 188-28-4-0-B12-CS and polynucleotide sequences encoding these biologically active fragments. "Biologically active fragments" are defined as those peptide or polypeptide fragments having at least one of the biological functions of the full length protein (e.g., the ability to transform cell lines in vitro.).

The invention also provides variants of the protein of SEQ ID NO: 311, encoded by clone 188-28-4-0-B12-CS. These variants have at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence encoded by clone 188-28-4-0-B12-CS. Variants according to the subject invention also have at

least one functional or structural characteristic of the protein encoded by clone 188-28-4-0-B12-CS. The invention also provides biologically active fragments of the variant proteins. Unless otherwise indicated, the methods disclosed herein can be practiced utilizing the protein encoded by clone 188-28-4-0-B12-CS, or clone 188-28-4-0-B12-CS, or variants thereof. Likewise, the methods of the subject invention can be practiced using biologically active fragments of the protein encoded by clone 188-28-4-0-B12-CS , clone 188-28-4-0-B12-CS, or variants of said biologically active fragments.

Because of the redundancy of the genetic code, a variety of different DNA sequences can encode the amino acid sequence provided by clone 188-28-4-0-B12-CS. It is well within the skill of a person trained in the art to create these alternative DNA sequences encoding proteins having the same, or essentially the same, amino acid sequence. These variant DNA sequences are, thus, within the scope of the subject invention. As used herein, reference to "essentially the same" sequence refers to sequences that have amino acid substitutions, deletions, additions, or insertions that do not materially affect biological activity. Fragments retaining one or more characteristic biological activity of the protein encoded by clone 188-28-4-0-B12-CS are also included in this definition.

"Recombinant nucleotide variants" are alternate polynucleotides which encode a particular protein. They can be synthesized, for example, by making use of the "redundancy" in the genetic code. Various codon substitutions, such as the silent changes which produce specific restriction sites or codon usage-specific mutations, can be introduced to optimize cloning into a plasmid or viral vector or expression in a particular prokaryotic or eukaryotic host system, respectively.

In one aspect of the subject invention, SEQ ID NO: 311, encoded by clone 188-28-4-0-B12-CS, and variants thereof, can be used to generate polyclonal or monoclonal antibodies. Both biologically active and immunogenic fragments of the amino acid sequence or variant proteins can be used to produce antibodies. Polyclonal and/or monoclonal antibodies can be made according to methods well known to the skilled artisan. Antibodies produced in accordance with subject invention can be used in a variety of detection assays known to those skilled in the art. Another aspect of this invention provides monoclonal and polyclonal antibodies which do not cross-react with known p120 proteins.

In one embodiment, the protein encoded by clone 188-28-4-0-B12-CS, variants of said protein, and biologically active fragments of the protein or said variants can be used as a nucleolar-fraction marker in nuclear fractionation studies or as a marker of pre-ribosomal particles, in methods well known to the skilled artisan.

In another embodiment, the protein encoded by clone 188-28-4-0-B12-CS, variants of said proteins, and biologically active fragments thereof, can be used as a proliferation marker in neoplastic cells. Alternatively, quantitative immunoassays can be used to assess the levels of the protein in resected cancerous and normal tissues. Alternatively, levels of the protein encoded by

clone 188-28-4-0-B12-CS can be compared between an individual and a "normal" control group as a prognostic indicator of malignancy. Further, the relationship between protein expression and cell proliferation can be assayed using cancer cell lines. Thus, the protein of the invention or part thereof can be used as a marker for proliferation in human cancer cells in vivo and in vitro. If the absence of expression of the protein of the invention on normal and benign tumors is confirmed, it could serve as a marker of malignant cancer cells. The proliferation rate of cancer cells can be also determined by quantitative analysis of the expression of the protein encoded by 188-28-4-0-B12-CS, or biologically active fragments thereof, according to methods described in Trere D. et al. (J. Pathol. 192:216-20, 2000).

The transforming activity of the protein of the invention can be assayed as described in Perlaky et al. (Anticancer Drug Des. 8:3-14, 1993). Thus, polynucleotides encoding the (188-28-4-0-B12-CS) protein can be used to induce transformation on NIH/3T3 cells in vitro. Alternatively, the polynucleotide encoding (188-28-4-0-B12-CS) can be used to provide antisense oligonucleotides useful in antisense therapeutic protocols according to methods known in the art.

Protein of SEQ ID NO: 406 (internal designation 174-32-4-0-F8-CS)

The 378-amino-acid-long protein of SEQ ID NO: 406 encoded by the cDNA of SEQ ID NO: 165 is expressed in tissues such as colon, prostate and salivary glands and overexpressed in colon and prostate. The C-terminus of the protein of the invention is homologous to the human retinoblastoma-binding protein, RbAp48 (Qian YW et al. (1993) Nature 364:648-652, GenBank accession number: X74262) and to its homologues conserved in other organisms including mouse (GenBank accession number: Q60972) and C. elegans (GenBank accession number: AF116530). The protein of the invention contains also two internal WD-repeat clusters (Prosite PS00678, amino acid positions 267 to 304 and positions 333 to 370, respectively), a structural motif involved in proteins interaction in signal transduction pathway and transcription regulation (Neer EJ et al. (1994) Nature 371:297-300; Neer EJ et al (1996) Cell 84:175-178).

The retinoblastoma protein (Rb) is the product of the retinoblastoma gene. Deletion or inactivation of both Rb alleles is essential in the formation of human retinoblastoma in both hereditary and sporadic forms (Benedict WF et al. (1983) Science 219: 973-975).

Loss-of-function mutations in the Rb gene is also found in many other tumor types, including osteosarcoma, breast carcinoma, small cell lung carcinoma, bladder carcinoma, prostate carcinoma and soft tissue sarcoma (Bookstein R et al. (1991) Crit. Rev. Oncog. 2:211-227). Introduction of the wild-type Rb gene into cultured retinoblastoma cells suppresses cells growth and their tumorigenicity in nude mice (Huang HJ et al (1988) 242:1536-1566). Expression of normal Rb protein in prostate carcinoma, osteosarcoma, breast carcinoma and bladder carcinoma cells also suppresses their neoplastic phenotype, thus establishing the Rb gene as a tumor suppressor (Reviewed by Weinberg RA (1991) Science 254:1138-1146).

It has been shown that the Rb gene product is a nuclear phosphoprotein that undergoes cyclic phosphorylation and dephosphorylation during cell cycling. Rb is unphosphorylated or "underphosphorylated" during early G1 phase, and become phosphorylated just before S phase, and remains phosphorylated until late mitosis. Injection of unphosphorylated Rb protein into cell during early G1 phase inhibits the entry into S phase, suggesting that some of the growth suppressor functions of Rb may be carried out by the underphosphorylated form of Rb (Goodrich et al. (1991) Cell 67:293-302; Hinds PW et al. (1992) Cell 70:993-1006). Rb protein not only regulates cell cycle, but is also involved in cell differentiation. For example, lens epithelial cells in Rb-deficient mouse fail to terminally differentiate and undergo apoptosis (Morgenbesser et al.(1994) Nature 371:72-74).

It has been demonstrated that Rb protein inhibits cellular growth and proliferation through interactions with multiple cellular proteins that interfere with these cellular protein's downstream actions. For example, Rb protein is able to form specific complexes with transcriptional factor E2F, which regulates the expression of a set of genes essential for the G1 to S phase transition (Nevin JR et al (1992) Science 258:424-429). The Rb protein restrains cell cycle progression by masking the E2F transactivation domain and by blocking the interaction of surrounding enhancer elements and basal transcription complex (Weintraub SJ et al (1995) Nature 375:812-815). Association of Rb and UBF, a ribosomal transcription factor, results in suppression of the synthesis of ribosomal RNA by RNA polymerase I (Cavanaugh LI et al (1995) Nature 374:177-180; Mancini M et al (1994) Proc.Natl.Acad.Sci.USA 91:418-422).

RbAp48 was first identified as a major protein from Hela cell that binds to a putative functional domain at the C-terminus of the Rb protein. Only unphosphorylated and hypophosphorylated forms of the Rb protein were coprecipitated with RbAp48. Like Rb protein, RbAp48 is a ubiquitously expressed nuclear protein that shares sequence homology with MSI1, a negative regulator of the Ras-cAMP pathway in the yeast Saccharomyces cerevisiae. Overexpression of RbAp48 can convert the yeast mutant strains from heat-shock sensitivity to heat-shock resistant, similar to the result obtained from MSI1 overexpression. Thus, the human RbAp48 is a functional homologue of MSI1 (Qian YW et al. (1993) Nature 364:648-652).

Rbap48 protein was later found to be the p48 subunit of mammalian chromatin assembly factor 1 (CAF-1) and to be present in histone deacetylase complex (Parthum MR et al (1996) Cell 87:85-94). CAF-1 from human cell nuclei consists of three subunits of p150, p60 and p48 and is involved in assembling of histone3 and histone4 onto nascent nucleosome structure during DNA replication in S phase (Kaufman FD et al (1995) Cell 81:1105-1114). Indeed, some transcriptional repressors function through the recruitment of the histone deacetylase complex (HDAC), the latter acts by acetylating or deacetylating the tail protruding from the core histones, thereby modulating the local structure of chromatin (Reviewed by Pazin MJ et al (1997) Cell 89:325-328). Rb protein recruits HDAC for binding to E2F to repress transcription (Brehm A et al (1998) Nature 391:597-601; Magnaghi-Jaulin L

et al (1998) Nature 391:601-604). It was also reported that the p48 subunit of chicken CAF-1 can bind to chicken HDAC in vitro through interaction of WD-40 repeats presented in both protein sequences (Ahmad A et al (1999) J.Biol.Chem. 274:16646-16653).

The WD-40 protein family is characterized by the repetition of a loosely conserved repeat of approximately 40 amino acids, each repeat being separated from each other by a Trp-Asp dipeptide sequence. The conserved core of this repeat, which usually ends with the amino acids Trp-Asp (WD), was first identified in the beta-subunit of the heterotrimeric GTP-binding protein, G-protein (Fong H et al. (1986) Proc.Natl.Acad.Sci.USA 83:2162-2166). Among the WD-40 proteins identified to date, none are enzymes, and all seem to have regulatory functions (Neer, E. J. et al. (1994) Nature 371:297-300). A number of WD repeat proteins have been localized to the nucleus and function in the repression of transcription. These include Tup1, Hir1, and Met30 in S. cerevisiae; SCON2 in Neurospora crassa; extra sex combs and Groucho in Drosophila; COP1 in Arabidopsis thaliana; and HIRA and the family of TLE proteins in humans. These WD-40 proteins turn off a wide variety of genes, including those involved in segmentation, sex determination, and neurogenesis (controlled by Groucho) and those involved in photomorphogenesis (controlled by COP1). All of these WD-40 containing proteins have been proposed to fold into propellers in which the internal beta-strands form a rigid skeleton that is fleshed out on the surface by specialized loops to which other proteins bind (Lambright DG et al (1996) Nature 379:311-319; Sondex J et al (1996) Nature 379:369-374).

Thus, discovery of new Rb-binding proteins is necessary to design methods of regulating cell growth and block tumorigenesis through the control of tumor suppressor proteins in their interaction with oncogene products and may provide new compositions which are useful in the diagnosis, prevention and treatment of cancer and developmental disorders.

It is believed that the protein of SEQ ID NO: 406 or part thereof plays a role in the control of gene expression, probably as a transcription repressor. The protein of the invention is thought to be able to bind to other proteins, preferably to nuclear proteins, more preferably to Rb. Preferred polypeptides of the invention are polypeptides comprising fragments of SEQ ID NO: 406 from position 159-373, 267-304 and 333-370. Other preferred polypeptides of the invention are polypeptides comprising fragments of SEQ ID NO: 406 having any of the biological activity described herein. The ability of the protein of the invention or part thereof to function as a transcription repressor may be assessed using techniques well known to those skilled in the art including those described previously (Weintraub SJ (1995) Nature 375:812-815; Qian YW (1995) J.Biol.Chem. 270:25507-25513). The ability of the protein of the invention or part thereof, especially fragments containing WD-repeats, to bind to other proteins may be assessed using techniques well known to those skilled in the art including those described herein. For example, the protein of the invention could be used as a "bait" protein in a yeast double hybridization system (e.g. Gal-4-based system from Clontech) to isolate and eventually to identify its interacting protein partner in vivo from a cDNA library. Alternatively, the protein of the invention or part thereof can be used either in a pure form or in a fusion form (linked to a reporter gene

product, such as alkaline phosphatase) to screen a phage cDNA expression library derived from selected tissues or cell types of a given organism (Scott et al (1990) Science 249:386-390; Lam et al (1992) Nature 354:82-84). Preferably, the binding ability of protein of the invention is tested in mammalian cell transfection experiments. When fused in-frame to a suitable peptide tag in expression vector, such as $[His]_6$ in the pRset expression plasmid vector (Invitrogen) and introduced into culture cells, the proteins that bind to the expressed fusion protein can be immunoprecipitated using anti-$[His]_6$ antibody. This approach can also be employed to confirm the findings obtained from either yeast double hybridization system or in vitro phage peptide library screening. In this case, the putative interacting partner protein will be fused to a distinct tag in a second expression vector and co-transfected into culture cells. True binding complex will be co-immunoprecipitated with the two different anti-tag antibodies. In a particular embodiment, an affinity chromatography method is carried out to identify the interacting protein partners in vitro from cell lysates as performed for the identification of the RbAp48 protein (Qian YW et al. (1993) Nature 364:648-652).

An embodiment of the present invention relates to methods of using the protein of the invention or part thereof, particularly polypeptides containing WD-motifs, or derivative thereof to identify and/or quantify binding proteins, preferably nuclear proteins, more preferably Rb, in a biological sample, and thus used in assays and diagnostic kits for the quantification of such binding proteins in bodily fluids, in tissue samples, and in mammalian cell cultures. Such assays may be particularly useful as diagnostic or prognostic tools in the detection and monitoring of a disorder linked to dysregulation of expression of a transcription regulator. Such assays may thus be very useful to asses the level of the tumor suppressor Rb in disorders including but not limited to developmental disorders, cancers such as retinoblastoma, prostate carcinoma, osteosarcoma, breast carcinoma and bladder carcinoma. The binding activity of the protein of the invention or part thereof may be assessed using any method familiar to those skilled in the art. Preferably, a defined quantity of the protein of the invention or part thereof is added to the sample under conditions allowing the formation of a complex between the protein of the invention or part thereof and the binding protein to be identified and/or quantified. Then, the presence of the complex and/or or the free protein of the invention or part thereof is assayed and eventually compared to a control using any of the techniques known by those skilled in the art.

Another embodiment of the present invention relates to compositions and methods using the protein of the invention or part thereof or derivative thereof to block gene transcription either in vitro or in vivo. In a preferred embodiment, the protein of the invention or part thereof or derivative thereof is added in an effective amount to an in vitro culture to inhibit gene expression and thus cell proliferation using molecular biology techniques known to those skilled in the art allowing the import of the protein from the extracellular medium to the cell's nucleus. In another embodiment, eukaryotic cells are genetically engineered in order to express the protein of the invention or part thereof under specific

conditions in order to prevent further proliferation of such cells upon demand such as infection, transformation, activation, differentiation, end of a production process.

A preferred embodiment of the invention relates to compositions or methods using SEQ ID NO: 406, SEQ ID NO: 165 or part thereof to diagnose, treat and/or prevent disorders including but not limited to disorders linked to dysregulation of gene transcription such as cancers and other disorders relating to abnormal cellular differentiation, proliferation, or degeneration, including hyperaldosteronism, hypocortisolism (Addison's disease), hyperthyroidism (Grave's disease), hypothyroidism, colorectal polyps, gastritis, gastric and duodenal ulcers, ulcerative colitis, and Crohn's disease; metabolic diseases such as obesity and a number of inflammatory diseases due to interleukin over-expression. For diagnostic and prognostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. For prevention and/or treatment purposes, the protein of the invention may be overexpressed using any of the gene therapy methods known to those skilled in the art including those described herein. For example, expression of the protein of the invention can be upregulated by infecting tumor cells with a retroviral or an adenoviral vector which expresses the desired protein at higher levels necessary for suppression of mutation in the Rb gene or in other oncogenic or tumor suppressor genes.

Another related embodiment relates to the use of SEQ ID NO: 406, SEQ ID NO: 165, its complement, or any part thereof to develop antagonists of the protein of the invention. These antagonists could be antisense oligonucleotides, triple helices, ribozymes, small molecules or antibodies, especially neutralizing antibodies binding to the WD-repeats of the invention, and may be used to treat disease and conditions caused by abnormally low transcription. These conditions include accelerated aging syndromes such as Cochayne's syndrome, Ataxia telangiectasia and Werner's syndrome as well as age-associated diseases as well as "early onset" forms of diseases associated with old age such as dementia and Parkinson's disease.

In another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably colon and prostate. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Protein of SEQ ID NO: 414 (internal designation 188-27-3-0-G1-CS)

The 389 amino-acid long protein of SEQ ID NO: 414, expressed in brain, fetal brain, placenta and testis, over-expressed in brain and encoded by the cDNA of SEQ ID NO: 173 is homologous to SIRTUIN-2 (SIRT2) (Trembl accession number: Q9Y6E9) and Silent Information

Regulator 2-like protein (SIR2L) (Trembl accession number: Q9UNT0) that belong to the Silent Information Regulator type 2 (SIR2) family. In addition, the protein of the invention presents the Pfam signature for members of the SIR2 family (amino acids 84-268). Furthermore, the protein of the invention displays two characteristic motifs highly conserved among all members of the SIR 2 family that have been shown to be essential in the SIR2 silencing function (Moira M. et al., Genetics, 154:1069-1083 (2000)). These motifs are from positions 84 to 98 and from positions 165 to 170 of the protein of SEQ ID NO: 414 and correspond to GAG(I/V)SxxxG(I/V)PDFERS and (Y/I)TQNID patterns respectively. The protein of SEQ ID NO: 414 also has conserved cysteines residues at positions 195, 200, 221 and 224, covering a domain thought to be either a DNA-binding zinc-finger motif (Prodom prediction PD002659, from positions 195 to 224) or an enzymatic domain (or an enzyme cofactor) (Moira M. et al., Genetics, 154:1069-1083 (2000)).

The cDNA of SEQ ID NO: 173 encoding the protein of the invention differs from the one encoding the SIRT2 protein by a supplementary exon between positions 147 to 195. This exon modifies the initiation codon of the protein and extends the ORF in its N-terminal part by 16 amino acids. Moreover, amino residues in positions 20 and 21 of the protein of the invention (respectively an alanine and a glutamine residue) are substituted from a glutamine and a tyrosine residue (positions 4 and 5) of the SIRT2 protein. Thus, the protein of the invention is a new isoform of SIR2 resulting from alternative splicing. The protein of the invention of SEQ ID NO: 414 is also 37 amino acids longer than the SIRL2 protein at its N-terminal end.

Regulation of gene expression by alterations in chromatin structure is a universal mechanism in eukaryotic cells, responsible for maintaining patterns of gene expression throughout the development of multicellular organisms. Silencing has been studied most extensively *in S. cerevisae* (yeast). Among the SIR genes, SIR2 is the most evolutionarily conserved, and a number of genes with homology to SIR2 have been identified (Frye R et al., Biochem. Biophys. Res. Commun., 273:793-798 (2000)). Presence of Homologues of SIR2 (*HSTs*) in organisms from bacteria to humans suggests that SIR2's silencing mechanism might be conserved. SIR2 was originally discovered to influence mating-type control in haploid cells by locus-specific transcriptional silencing. It has also been suggested that SIR2 and its homologs play additional roles in suppression of recombination, chromosomal stability, metabolic regulation, meiosis, and aging (for a review: see Gartenberg, Curr. Opin. Microbiol. 3:132-137 (2000)).

Proteins of the SIR2 family are also thought to be either enzymes or enzyme cofactors. First, Landry and collaborators have shown that members of SIR2 family catalyze histone deacetylation in a reaction that requires NAD, thereby distinguishing them from previously characterized deacetylases. This enzyme is active on histone substrates that have been acetylated by both chromatin assembly-linked and transcription related acetyltransferases (Landry et al., Proc. Natl. Acad. Sci. 97:5807-5811 (2000)). Discovery of an intrinsic deacetylation activity for the conserved SIR2 family provides a mechanism for modifying histones and other proteins to regulate

transcription and diverse biological process. Secondly, the study of a human SIR2 family member (hSirT2) was found to have a mono-ADP ribosylation activity *in vitro* (Frye R et al., Biochem. Biophys. Res. Commun., 260: 273-279 (1999)). Among potential substrates for mono-ADP ribosylation are histones and RNA Pol I, modification of which correlates with enhanced rDNA transcription.

It is believed that the protein of SEQ ID NO: 414 or part thereof plays a role in gene silencing, suppression of recombination, chromosomal stability, metabolic regulation, meiosis, and aging, probably as a member of the SIR2 protein family. Particularly, the protein of the invention may deacetylate substrates, preferably acetylated histones and acetyltransferases, either directly or indirectly as enzyme cofactors. Particularly, the protein of the invention may have a ribosylation activity, preferably a mono-ADP ribosylation activity, preferably on histones and RNA Pol I substrates, either directly or indirectly as enzyme cofactors. Additionally, the protein of the invention may be a DNA binding protein. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 414 from positions 1 to 16, 84 to 98, 165 to 170, 195 to 224, and 84-268 as well as fragments of SEQ ID NO: 414 containing at least one cysteine residues located in positions 195, 200, 221 or 224 of SEQ ID NO: 414. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 414 having any of the biological activities described herein. The deacetylation activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in Laundry et al., supra. The ribosylation activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in Frye et al, (1999). The nucleic acid binding activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in US patent 6,013,453.

The invention relates to methods and compositions using the protein of the invention or part thereof to silence gene expression. In a preferred embodiment, the protein of the invention or part thereof or derivative thereof is added in an effective amount to an in vitro culture to inhibit gene expression and thus cell proliferation using molecular biology techniques known to those skilled in the art allowing the import of the protein from the extracellular medium to the cell's nucleus. In another embodiment, eukaryotic cells are genetically engineered in order to express the protein of the invention or part thereof under specific conditions in order to prevent further proliferation of such cells upon demand such as infection, transformation, end of a production process, differentiation, etc...

The invention relates to methods and compositions using the protein of the invention or part thereof to deacetylate substrates, alone or in combination with other substances. Such substrates are acetylated substrates, preferably acetylated histones and acetyltransferases. For example, the protein of the invention or part thereof is added to a sample containing the substrate(s) in conditions allowing deacetylation, and allowed to catalyze the deacetylation of the substrate(s). In a preferred

embodiment, the deacetylation is carried out using a standard assay such as those described in Laundry et al, supra. Deacetylated histones obtained by this method may be mixed with purified naked DNA (plasmid preparations for example) in order to reconstitute chromatine-like structures *in vitro*. Such structures are of great interest in the study of enzymatic factors involved in transcription and replication.

Another embodiment of the present invention relates to composition and methods of using the protein of the invention or part thereof to develop assays for *in vitro* screening of inhibitors directed against the encoded deacetylase activity using any technique known to those skilled in the art including those described herein. Such deacetylase inhibitors are of great potential as new drugs due to their ability to influence transcriptional regulation and to induce apoptosis or differentiation in cancer cells. Preferably, the protein of the invention, expressed in prokaryotic or eukaryotic systems according to methods known to those skilled in the art, may be mixed *in vitro* with a simple fluorescent substrate like an aminocoumarin derivative of an acetylated lysine, and different putative inhibitors. The coumarin derivative is then quantitated using a reverse-phase HPLC-system with a fluorescence detector. Such an approach has been previously developed by Hoffmann and collaborators (Hoffmann et al., Nucl. Acids Res. 27:2057-2058 (1999); Hoffmann et al., Pharmazie 55:601-606 (2000)).

The invention relates to methods and compositions using the protein of the invention or part thereof to bind to nucleic acids, preferably DNA, alone or in combination with other substances. For example, the protein of the invention or part thereof is added to a sample containing nucleic acid in conditions allowing binding, and allowed to bind to nucleic acids. In a preferred embodiment, the protein of the invention or part thereof may be used to purify nucleic acids such as restriction fragments. In another preferred embodiment, the protein of the invention or part thereof may be used to visualize nucleic acids when the polypeptide is linked to an appropriate fusion partner, or is detected by probing with an antibody. Alternatively, the protein of the invention or part thereof may be bound to a chromatographic support, either alone or in combination with other DNA binding proteins, using techniques well known in the art, to form an affinity chromatography column. A sample containing nucleic acids to purify is run through the column. Immobilizing the protein of the invention or part thereof on a support advantageous is particularly for those embodiments in which the method is to be practiced on a commercial scale. This immobilization facilitates the removal of the protein from the batch of product and subsequent reuse of the protein. Immobilization of the protein of the invention or part thereof can be accomplished, for example, by inserting a cellulose-binding domain in the protein. One of skill in the art will understand that other methods of immobilization could also be used and are described in the available literature.

Still another embodiment of the present invention relates to composition and methods of using the protein of the invention or part thereof to identify genes or regions of the human genome silenced by the protein of the invention or part thereof. Genomic DNA derived from patients with

pathologies such as cancer and metabolic disorders, or from elderly people may be compared to those extracted from respective controls for their ability to bind the protein of the invention. As described previously, the protein of SEQ ID NO: 414 displays a putative zinc finger domain susceptible to bind DNA sequences near regions or genes to silence. The protein of the invention or part thereof may be bound to a chromatographic support, using techniques well known in the art, to form an affinity chromatography column. A sample containing a mixture of human genomic DNA digested by endorestriction enzymes is run through the column. After extensive washings the bound DNA is eluted and further subcloned in classical cloning vectors known to those skilled in the art. Immobilizing the protein of the invention or part thereof on a support is particularly advantageous for those embodiments in which the method has to be practiced routinely. This immobilization facilitates the removal of DNAs from the batch of resin coupled protein after binding, and allows subsequent re-use of the protein. Immobilization of the protein of the invention or part thereof can be accomplished, for example, by inserting any matrix-binding domain in the protein according to methods known to those skilled in the art. The resulting fusion product including the protein of the invention or part thereof is then covalently, or by any other means, bound to a protein, carbohydrate or matrix (such as gold, "Sephadex" particles, and polymeric surfaces).

Another embodiment of the invention relates to methods of preparing antibodies directed against the protein of the invention or part thereof. Such antibodies may be used in co-immunoprecipitation procedures that enrich for chromatin fragments containing binding sites for the protein of the invention. This method may identify genes or regions of the human genome silenced by the deacetylase activity of the protein of the invention. Preferably, in samples containing fragments of native chromatin, antibodies directed against 414 and coupled to protein A or protein G sepharose beads are added to the mixture. Immunoprecipitation conditions are those known to those skilled in the art. After washings DNA fragments co-precipitated with 414 are extracted and further subcloned in routinely used cloning vectors. These DNA fragments are either sequenced and/or used as probes to screen genomic libraries. This procedure is very similar to the one used by Gould and collaborators to enriche for embryonic chromatin fragments containing sites for the homeotic Ubx protein (Gould et al., Nature 348:308-312 (1990)).

A preferred embodiment of the invention relates to compositions or methods using SEQ ID NO: 414, SEQ ID NO: 173 or part thereof to diagnose, treat and/or prevent develop disorders caused by the expression of "disease causing" genes. The number of pathologies and conditions that could be treated by the protein of the invention is potentially huge and unlimited. Favored disorders linked to dysregulation of gene transcription such as cancer and other disorders relating to abnormal cellular differentiation, proliferation, or degeneration, including hyperaldosteronism, hypocortisolism (Addison's disease), hyperthyroidism (Grave's disease), hypothyroidism, colorectal polyps, gastritis, gastric and duodenal ulcers, ulcerative colitis, and Crohn's disease; viral infection especially HIV and viral hepatitis (i.e. expression of viral proteins), metabolic diseases such as

obesity and a number of inflammatory diseases due to interleukin over-expression. For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods described herein and compared to the expression in control individuals. For prevention and/or treatment purposes, the protein of the invention may be overexpressed using any of the gene therapy methods known to those skilled in the art including those described herein. For example, switching off "disease" genes may be achieved by, for example, directly targeting the protein of the invention or part thereof to the genes (such as oncogenes in cancers) that are over-expressed in order to silence their expression. This could be achieved by making a "chimera" protein in which the putative zinc-binding domain is replaced by a sequence known to bind to, or near the over-expressed gene as explained elsewhere in the application. Fusion proteins containing both the deacetylase activity and the specific DNA binding domain are obtained by methods of molecular biology well known to those skilled in the art. The corresponding eukaryotic expression vectors may be use in gene therapy in the cases of cancer, metabolic disorders, aging and any disorder where a gene is over-expressed. Such recombinant cDNA may be introduced in the well known adenoviral vectors used in cancer therapy (for a recent review on the use of replicative adenoviruses for cancer therapy : Alemany et al., Nat. Biotechnol. 18:723-727 (2000)).

Another related embodiment relates to the use of SEQ ID NO: 414, SEQ ID NO: 173, its complement, or any part thereof to develop antagonists of the protein of the invention and of the SIR complex. These antagonists could be antisense oligonucleotides, triple helices, ribozymes, small molecules or antibodies and may be used to treat disease and conditions caused by abnormal gene silencing. These conditions include accelerated aging syndromes such as Cochayne's syndrome, Ataxia telangiectasia and Werner's syndrome as well as age-associated diseases as well as "early onset" forms of diseases associated with old age such as dementia and Parkinson's disease.

In another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably brain tissues. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Protein of SEQ ID NO:298 (182-1-2-0-D12-CS)

The protein of SEQ ID NO:298, encoded by the cDNA of SEQ ID NO:57, is homologous to proteins of the fibroblast growth factor family (FGF). Specifically the amino acid sequence of SEQ

ID NO:298 is identical to the recently described FGF-23. The protein of the invention is strongly expressed in the fetal liver.

The protein of the invention presents the pfam signature for fibroblast growth factors (positions 48 to 129). High resolution X-ray structures of crystals of both FGF-1 and FGF-2 have been reported and reveal a "beta trefoil" topology, comprising 12 strands linked to form a three-fold symmetrical structure made up of four-stranded antiparallel beta sheet (see Faham S. et al. - Curr Opin Struct Biol. - 1998, 8(5): p578-586). On the basis of sequence conservation, it seems very likely that all members of the FGF family have related 3-dimesional structures. Preferred polypeptides of the invention are those that comprise amino acids 39 to 45; 51 to 56; 60 to 64; 71 to 77; 82 to 87; 92 to 97; 101 to 105; 113 to 119; 124 to 130; 142 to 147; 151 to 155 and/or 167 to 172, which by homology with other members of the FGF family make up the 12 beta pleated sheets characteristic of the FGF family (White K. et al. - Nat Genet. - 2000; 26(3): pp. 345-348). Furthermore, as within these regions a number of amino acids from SEQ ID NO:298 are conserved in over 80% of human FGFs (after sequence alignment), the most preferred polypeptides of the invention comprise amino acids 42, 53, 63, 83, 85, 87, 93, 96, 101, 113, 115, 124, 127 and/or 129. Other preferred polypeptides of the invention are any fragment of SEQ ID NO:298 having any of the biological activities described herein.

Cytokines are a heterogeneous group of polypeptide mediators associated with numerous functions, including immune system and inflammatory responses. The cytokine families include, but are not limited to, Interleukins, Chemokines, Tumor necrosis factors, Interferons, Colony stimulating factors, Neurotrophins, neuropoietins and growth factors (of which the FGF family is a member). Fibroblast growth factors (FGFs) were first characterized, in the mid 1970s, as mitogens of cultured fibroblasts. Since then more then 20 different FGFs have been identified. Fibroblast growth factors belong to a family of proteins called growth factors (other members of this family include EGF, PDGF, TGFs and ECGF). The biological effects of FGFs are mediated by association with 3 biochemically distinct partners: heparan sulfate proteoglycans, a low affinity transmembrane FGF-binding protein and high-affinity transmembrane FGF receptors of the tyrosine-kinase class. Transfection and reconstruction experiments have shown that intracellular signal transduction is triggered by activation of FGF receptor kinase activity. Activation is brought about by receptor oligomerization, which is mediated by the association of heparan sulfate proteoglycans with the ligand (FGF) and of the ligand with the receptor itself (Faham S. et al. - Curr Opin Struct Biol. - 1998, 8(5): p578-586). Longer heparin-derived oligosaccharides generally exhibit tighter binding to FGF. The relationship between heparin length, biological activity and FGF binding has been extensively studied and there is general agreement that longer heparin oligosaccharides tend to be more biologically active. FGFs are members of a family with a broad range of biological activities involving cell growth and differentiation (including angiogenesis, morphogenesis and wound healing); cell survival, replication, adhesion and mobility. FGFs have been found to be potent

growth factors for a number of cell types including, but not limited to fibroblasts, endothelial cells, smooth muscle cells, keratinocytes, osteoblasts and neurons.

Clearly, FGF biology is potentially very complex, involving multiple ligands, receptors and cofactors, each expressed with different spatial and temporal patterns and distinct kinetics in the course of normal development. Considerable efforts have been expended on the creation of different types of animal models for the analysis of FGF function in vivo. These studies clearly indicate that FGF signaling is involved in a number of different processes at different stages of development and is critical in early developmental stages (FGF-4 and FGF receptor 1 homozygous null mutations both cause early lethality). FGF signaling has been found to be required for both branching morphogenesis of the lung and the establishment of the normal program of keratinocyte differentiation in the skin. FGF signaling has also been found to be involved in both the initial induction and sustained outgrowth of the limb bud during early limb development. Perhaps the most impressive illustration of this function of FGF signaling is the ability to induce supernumerary limb development in the chick by local application of FGF-soaked beads (Cohn M, et al. - Cell – 1995; 80: p739-746), thus indicating that a least some FGF-dependent processes are regulated by accessibility of an FGF ligand. FGFs are also capable of stimulating migration and differentiation of hepatic precursors.

Recently mutations in the FGF-23 gene were found to be associated with autosomal dominant hypophosphataemic rickets (ADHR), a genetically transmitted disease characterized by low serum phosphorus concentrations, rickets, osteomalacia, lower extremities deformation, short stature, bone pain and dental abscesses. It seems very likely that FGF signaling functions are involved in numerous aspects of morphogenesis, differentiation and other essential cellular mechanisms, and are thus likely involved in any of a large number of diseases and conditions associated with these processes.

Thus, it is believed that the protein of SEQ ID NO:298 is a member of the fibroblast growth factor family, and is thus involved in a large number of cellular and organismal proesses, including, but not limited to, cell growth and differentiation, angiogenesis, morphogenesis, wound healing, cell survival, replication, adhesion and mobility.

One embodiment of the present invention relates to the use of the present polypeptides and polynucleotides to identify liver, heart, thyroid and parathyroid tissues, or cells derived from these tissues, since the protein of the invention is expressed therein (White K. et al. - Nat Genet. - 2000; 26(3): p345-348). Such detection of cells expressing the protein can be carried out in any of a number of ways, including the use of specific antibodies or antiserum generated against the protein using standard methods, as well as using polynucleotide probes specific for nucleic acids encoding the protein of the invention.

In another embodiment, the protein of the invention or part thereof can be used as a mitogen to stimulate the growth of a number of different cells types including, but not limited to, fibroblasts,

muscle cells, osteoblasts, keratinocytes and hepatocytes. The growth of cells can be stimulated in vitro, for example to promote the growth of cells cultured for the synthesis of recombinant proteins, or for ex vivo gene therapy applications. Another preferred application of this technique relates to the use of the protein of the invention or part thereof to generate in vitro tissues and organs including, but not limited to, skin, cartilage, and bone for transplants and grafts (Lancet – 1981, 1(8211):75-8)).

Another preferred embodiment of the invention relates to the use of the invention or part thereof to treat damaged tissues and organs. Members of the FGF family have been shown to induce the differentiation and growth of a number of different cell types. Thus the protein of the invention can be administered to treat pathologies and conditions that result from damage to cells, tissues or organs. These pathologies and conditions include but are not limited to bone fractures and bone defects Kimoto et al. – J Dent Res – 1998, 77(12): p1965-1969) (Solheim E – Int Orthop – 1998, 22(6), damage due to wounds (such as lesions of the skin and ulcers) (Debus E. – Zentralbl Chir – 2000, 125 (supple 1): p49-55) (Szabo S. – Aliment Pharmacol Ther – 2000, 14(Suppl 1): p33-43), tissue damage due to ischemia (for example, in the brain and heart) (Simons M. – Circulation – 2000, 102(11): pE73-E86), cardiovascular diseases such as thrombosis and atherosclerosis (Bauters C. – Drugs – 1999, 58 (Spec No1): p11-15) and neurodegenerative diseases due to neuronal loss such as Parkinson's disease or Alzheimer's disease (Ebadi M – Neurochem Int – 1997, 30(4-5): p347-374) (Brundin P. – Cell Transplant – 2000, 9(2): p179-195).

In a most preferred embodiment, the polypeptides or polynucleotides of the invention can be used to diagnose, treat, or prevent disorders resulting from non-functional and/or mutated FGFs, such as Autosomal dominant hypophosphataemic rickets, which is associated with mutation of certain amino acids of FGF-23 (White K. et al. - Nat Genet. - 2000; 26(3): p345-348, which is hereby incorporated by reference in its entity. Such disorders can be treated, for example, by administering a therapeutically effective amount of the protein of the invention or a polynucleotide sequence encoding the protein to a patient suffering from the disorder. Similarly, SEQ ID NO:298 or SEQ ID NO:57 or any part thereof can be used to develop diagnostic kits in order to diagnose, prevent and/or treat any other disease associated with FGF, for example pathologies associated with FGF overexpression.

In yet another embodiment, the protein of the invention or part thereof can be used to develop antogonists of FGF and/or FGF receptors in order to treat disorders associated with an over-activation of FGF pathways (for example, due to over production of FGF or overstimulation of FGF receptors). This is particularly true for pathologies such as cancers where some tumors secrete large quantities of FGF, such as prostate and breast cancers. Furthermore FGF antagonists can be useful in inhibiting tumor angiogenesis, which is an essential step in tumor growth. In the same way SEQ ID NO:**57** or any part thereof could be used to generate antisense oligonucleotides. Antisense oligonucleotides block complementary mRNA, thus inhibiting the synthesis of the

protein encoded by the mRNA. These oligonucleotides can be used in in vivo or ex vivo treatment of the diseases caused or aggravated by overexpression of FGFs.

Protein of SEQ ID No: 396 (internal designation: 160-12-1-0-D10-CS)

The protein of SEQ ID No: 396 encoded by the cDNA of SEQ ID No: 155, overexpressed in brain and fetal brain, shows homology to memebers of the transmembrane 4 super family of proteins (TM4SF). Tthe protein of the invention displays signatures characteristic of this family, namely the pfam domain from positions 66 to 273, the Prosite motif from positions 112 to 134 as well as the emotif domains from positions 108 to 127, 108 to 146, 129 to 150, 128 to 154, and 247 to 274. In addition, the protein of the invention has several predicted transmembrane domains: 103 to 123, 130 to 150 and 245 to 265, with an additional predicted domain with lower certainty from positions 61 to 81. The protein of the invention has significant homology to a TM4SF member, the integral membrane CD81 antigen also known as TAPA1 (Target of Antiproliferative Antibody), except for its N-terminus. The transmembrane domains of the protein of the invention matches those described for CD81.

Members of the tetraspan family of proteins are associated with adhesion molecules and translate adhesive events into a regulation of cellular behaviour. TAPA-1 is a widely expressed protein found to influence adhesion, morphology, activation, proliferation and differentiation of B, T and other cells. TAPA-1 has two long hydrophilic domains of the molecule which are extracellular and located between four TM (Transmembrane region, TM1-4). The region between TM2 and TM3 is highly conserved in all tetraspanins. The protein is highly hydrophobic and contains a potential N-myristoylation site. TAPA1 functions by forming a complex on the cell surface and the antigenic epitope of the human TAPA1 is contained within a subregion of the second extracellular domain of the protein. Cell-surface expression of TAPA1 can be down-modulated by binding of antibodies (Levy 1991, J Biol Chem Aug 5;266(22):14597-602).

Mice lacking CD81 (not expressing TAPA1) have impaired antibody responses to protein antigens. This defect is specific to antigens that preferentially stimulate a T helper 2 response and is only seen with T cell-dependent antigens. Absence of CD81 on B cells is sufficient to cause the defect. Antigen-specific interleukin (IL) 4 production is greatly reduced in the spleen and lymph nodes of CD81-null mice compared with heterozygous littermates. The expression of CD81 on B cells is critical for inducing optimal IL-4 and antibody production during T helper 2 responses. CD81 is likely to have a greater role in the control of immune responses than in the development of immune cells (Maecker (1997) J Exp Med 1997 Apr 21;185(8):1505-10). CD81 on B cells has the capacity to promote IL-4 secretion from T cells. Costimulatory proteins such as B7-1 and B7-2 have been shown in some systems to have differential effects on cytokine secretion by T cells. CD81 on B cells, can control cytokine production by T cells. TAPA-1 has been implicated to play an important role in the regulation of lymphoma cell growth.

TAPA-1 is highly expressed in many neurons of the brainstem. TAPA is found in all glial cells, and the level of this protein correlates with their maturation (Sullivan et al., 1998, J Comp Neurol 1998 Jul 6;396(3):366-80). This protein is expressed by ependyma, choroid plexus, astrocytes, and oligodendrocytes. TAPA1 is dramatically upregulated during early postnatal development, at the time of glial birth and maturation. At embryonic day 18, the levels of TAPA are low, with most of the immunoreaction product being associated with the ependyma, choroid plexus, and the glia limitans. The amount of TAPA expressed in the brain increases with brain development, and at postnatal day 14 the protein levels approach those of the adult. This increase in the levels of TAPA at postnatal day 14 is due to upregulation in the gray matter and white matter. TAPA has been associated with reactive gliosis and the glial scar. The spatiotemporal expression pattern of CD81 by reactive microglia and astrocytes indicates that CD81 is involved in the glial response to spinal cord injury. It is suggested that the upregulation of TAPA is an integral component of glial scar formation (Peduzzi et al, Exp Neurol. 1999 Dec;160(2):460-8).

The levels of TAPA-1 are low in metastatic prostate tumors, expression of this protein in these cells appears to suppress metastatic behavior (Dong et al., 1995 Science. 12;268(5212):884-6.). Bivalent antibodies directed against these proteins can be used to enhance adhesion of different cell types: pre-B cells (Masellis-Smith and Shaw (1994) J Immunol 1994 Mar 15;152(6):2768-77), endothelial cells (Forsyth, 1991 Immunology Feb;72(2):292-6), and tumor cell mobility and invasiveness (Miyake et al., 1991 J Exp Med. Dec 1;174(6):1347-54.). In the nervous system, the migratory behavior of Schwann cells over biologically relevant substrates can be enhanced with the application of antibodies directed against certain TM (Anton et al (1995) J Neurosci. Jan;15(1 Pt 2):584-95). Antibodies directed against TAPA-1 depress the mitotic activity and induce an increase in cellular adhesion (Oren et al, 1990 Mol Cell Biol Aug;10(8):4007-15).

It is believed that the protein of SEQ ID NO: 396 or part thereof plays a role in cell adhesion, motility, metastasis, cell activation, signal transduction and the immune response, probably as a member of the TM4SF family. As a member of the tetraspanin family of proteins, the protein of SEQ ID No: 396 or part there of is believed to mediate cellular interaction in lymphoid cells as well as non-hematolymphoid tissue to affect cell adhesion and migration, alter cell morphology and the activation state of a cell. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 396 from positions 66 to 273, 112 to 134, 108 to 127, 108 to 146, 129 to 150, 128 to 154, and 247 to 274. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 396 having any of the biological activity described herein. The activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those describing a functional tissue assay used to define surface antigens regulating astrocyte growth (Eldon et al, 1996, J Neurosci, 16(17):5478); cellular function assays determining the involvment of the protein in signal transduction and cell adhesion

in the immune system (Levy et al, 1998 Ann. Rev. Immunol. 16:89-109, Virtaneva et al, 1994 Immunogenetics 39: 329-334).

An embodiment of the present invention relates to methods of using the protein of the invention or part thereof to identify and/or quantify membrane proteins, preferably integrins, lineage specific molecules, tetraspanins, and antibodies, in a biological sample, and thus used in assays and diagnostic kits for the quantification of such membrane proteins in tissue sample, and in mammalian cell cultures. The binding activity of the protein of the invention or part thereof may be assessed using the assay described in Shoshana et al, 1998, Annu. Rev. Immunol 16: 89-109; Maecker et al, 1998 PNAS 95: 2458-2462; Geisert et al, 1996, J of Neuroscience 16(17): 5478-5487 or any other method familiar to those skilled in the art. Preferably, a defined quantity of the protein of the invention or part thereof is added to the sample under conditions allowing the formation of a complex between the protein of the invention or part thereof and the membrane protein to be identified and/or quantified. Then, the presence of the complex and/or or the free protein of the invention or part thereof is assayed and eventually compared to a control using any of the techniques known by those skilled in the art.

In another embodiment, the invention relates to compositions and methods using the protein of the invention or part thereof to stimulate cell proliferation, preferably proliferation of lymphoid cells both in vitro and in vivo. For example, soluble forms of the protein of the invention or part thereof may be added to cell culture medium in an amount effective to stimulate cell proliferation.

In another embodiment, the invention relates to compositions and methods using the protein of the invention or part thereof or derivative thereof to stimulate antibody production either in vitro or in vivo. In a preferred embodiment, the protein of the invention or part thereof or derivative thereof may be added in an effective amount to stimulate antibody production to an in vitro culture of antibody-producing cells, such as hybridomas. In another preferred embodiment, the protein of the invention or part thereof or derivative thereof may be injected into an animal in order to increase the animal's antibody production to a protein of interest in the case of production of polyclonal antibodies.

In still another embodiment, the invention relates to compositions and methods using the protein of the invention or part thereof or derivative thereof to decrease cell adhesion either in vitro or in vivo. In a preferred embodiment, the protein of the invention or part thereof or derivative thereof may be added in an effective amount to prevent and/or inhibit cell adhesion to an in vitro culture of adherent cells in order to recover those adherent cells.

In still another embodiment, , the invention relates to compositions and methods using the protein of the invention or part thereof to treat and/or prevent cell-proliferative disorders, such as cancers,, via the prevention of metastatispreferably brain cancer, and disorders characterized by depressed immune response such as autoimmune diseases AIDS, allergy, type1 diabetes, systemic lupus erythematosus, chronic rheumatoid arthritis, juvenile rheumatoid arthritis, Sjogren's

syndrome, systemic scleriasis, mixed connective tissue disease and dermatomyositis, Hashimoto's disease, primary myxedema, thyrotoxia, pernicious anemia, ulcerative colitis, autoimmune atrophic gastritis, idiopathic Addison's disease, male infertility, Goodpasture's syndrome, acute progressive glomerular nephritis, myasthenia gravis, multiple myositis, pemphigus vulgaris, bullous pemphigoid, sympathetic ophthalmia, multiple sclerosis, autoimmune hemolytic anemia, idiopathic thrombocytopenic purpura, postmyocardial infarction syndrome, rheumatic fever, lupoid hepatitis, primary biliary cirrhosis, Behcet's syndrome and Crest's syndrome, via the stimulation of antibody production and IL-4 secretion.

In another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably from brain and fetal brain origin. For example, the protein of the invention or part may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry or any other technique known to those skilled in the art.

Protein of SEQ ID No: 296 (internal designation 181-3-3-0-B8-CS)

The protein of SEQ ID NO: 296 encoded by the cDNA of SEQ ID NO:55, overexpressed in fetal liver, is homologous to the whole domain IV-4 and IV-5 of the basement membrane-specific heparan sulfate proteoglycan core protein (perlecan), well conserved among C.elegans, mice and human (accession numbers Q06561, Q05793 and P98160 respectively). The 247-amino-acid-long protein of the invention, displays two putative hydrophobic stretches from positions 44 to 64 and 219 to 239 and a putative immunoglobulin domain from positions 141 to 197, homolog to the Ig domain 4 of the Ig repeat structure of domain IV of perlecan proteins. The protein of the invention displays also a putative secreted signal peptide from positions 6 to 21.

Basement membranes are specialized regions of extracellular matrix (ECM) containing a large number of different components, including laminin, collagen, nidogen and heparan sulfate proteoglycans (for a review see Bernfield at al., Annu. Rev. Biochem. 68:729-777 (1999)). Perlecan, a major basement membrane, plays important roles in many fundamental development and regenerative processes, including cell cohesion, adhesion and migration, signal transduction, and even gene regulation (Martin and Timpl, Annu. Rev. Cell Biol. 3:57-85 (1987)). The cDNA sequence of perlecan encodes a large core protein consisting of five structural domains, referred from I to V, with distinct motifs such as SEA modules (domain I), LDL class A modules (domain II), cysteine-rich LE modules (domain III), LAMB modules (domain V). Domain IV consists of Ig-like repeats (14 in mice, 21 in human perlecan) similar to those of neural cell adhesion molecules (N-CAMs). Glycosaminoglycan chains are mostly linked to domain I and have been shown to

participate with the core protein in its differential expression in tissues and development stages (Perrimon and Bernfield, Nature 404:725-728 (2000)).

The N-terminal fragment IV containing Ig modules from 1 to 8 show high-affinity binding to the two known nidogen isoforms, laminin1-nidogen1 complex (LN) and binding to heparin at physiological ionic strength (Hopf et al., Eur. J. Biochem. 259:917-925(1999)). An alteration study of the C. elegans perlecan show in vivo that mutations in Ig-like modules 3 and 4 induce a lethal phenotype by inhibiting the spatial distribution of the splice variants. Mutations inducing deletions in other Ig-like modules of perlecan domain IV does neither affect the isoform expression nor the spatial distribution, suggesting a crucial role of Ig-like modules 3 and 4 in muscle assembly and development stages (Mullen et al., Mol. Biol Cell 10:3205-3221 (1999)).

Several studies have shown the large presence of distinct perlecan isoforms through regulated alternative splicing in C. elegans (Rogalski et al, Genes Dev. 7:1471-1484 (1993), Rogalski et al, Genetics139:159-169(1995)). Although splice variants have not yet been shown in human, Ig-like modules are encoded by multiple exons compatible with different combinatorial possibilities of expression (Cohen et al., P.N.A.S. 90:10404-10408 (1993)). Alternative splicing within Domain IV is associated with temporal and spatial differences in isoform expression. A subset of C.elegans isoforms are associated with body-wall muscles during embryiogenesis and are required for nematode myofilament lattice assembly, which is very similar to assembly of focal adhesions in mammalian cell culture (Moerman and Fire, CSH labo. Press (1997)).

Basement membrane-like structure containing perlecan, collagen IV, laminin also plays a major role during liver differentiation by interacting with immature hepatocytes (Am. J. Path.142:199-208 (1993)).

Perlecan have been implicated in a number of processes and diseases resulting from the alteration of its structure including glomerular filtration deficiencies such as proteinuria, diabetic glomerulopathies, nephrotic syndromes, Denys-Drash syndromes (Groffen at al., Nephrol. Dial. Transplant 14:2119-2129 (1999)), mitogenesis and angiogenesis diseases (Aviezer et al., Cell 79:1005-1013 (1994)), inflammation and tissue repair, ocular and skeletal defect syndromes, microbial pathogenesis through invasion. Perlecan core protein has binding epitopes for the basement membrane proteins nidogen-1, nidogen-2, and fibulin-2, as well as for Alzheimer's beta-amyloid protein (Snow et al.,Arch. Biochem. Biophys. 320:84-95 (1995)) and platelet growth factor.

It is believed that protein of SEQ ID NO: 296 or part thereof is a membrane basement-like protein, preferably a human isoform of the perlecan protein. Thus, the protein of the invention plays an important role in membrane integrity and interactions with other basement proteins and particularly with nidogen-1 and 2, LN complexes and heparin coumpounds. Besides, the protein of the invention is thought to participate in the interactions with cellular receptors such as integrins, with cytokine release and proteolysis, with regulation of angiogenesis , wound healing and tumor

invasion. Being overexpressed in the fetal liver, the protein of the invention is thought to participate in the differentiation, migration and adhesion of hepatocytes trought its spatial and temporal expression during embryogenesis. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO: 296 from positions 141 to 197. Other preferred polypeptides of the invention are fragments of SEQ ID NO: 296 having any of the biological activity described herein. The activity of the protein of the invention or part thereof may be assayed using any of the assays known to those skilled in the art including those described in Hopf et al., Euro. J. Biochem. 259:917-925(1999) for binding assays with other membrane proteins and in Rescan et al., Am. J. Path. 142:199-208 (1993) for protein assays (Immunochemistry and ELISA).

In one embodimen, the invention relates to methods and compositions using the protein of the invention or part thereof as a new marker protein to selectively identify embryogenic stages, preferably in liver tissues. For example, the protein of the invention or part thereof may be detected using specific antibodies able to bind to the protein using any technique known to those skilled in the art. Such tissue-specific antibodies may then be used to identify embryogenic cells with dysregulated membrane components such as in differentiated tumor cells or to differentiate different cell types in a tissue cross-section using immunochemistry. For example, the amount of the protein of the invention in embryogenic cells reflecting the characterized overexpression activity is measured and compared to that of a normal cell using a specific antibody detected by fluorescence (FACS, confocal microscopy,...) or any other detection methods skilled in the art.

In another embodiment the invention relates to methods and compositions using the protein of the invention or part thereof for the diagnosis of a disorder associated with overexpression of the protein of the invention, preferably but not limited to perlecan associated tumors such as human melanoma, proliferative diseases, glomerular filtration deficiencies such as proteinuria, diabetic glomerulopathies, nephrotic syndromes, Denys-Drash syndromes, mitogenesis and angiogenesis diseases, inflammation and tissue repair, ocular and skeletal defect syndromes, microbial pathogenesis through invasion. The expression of the protein of the invention could be investigated using any methods well known to those skilled in the art, including Northern blotting, RT-PCR or immunoblotting using specific antibodies binding to the protein of the invention.

In still another embodiment the protein of the invention or part thereof could be used as a mitogen to stimulate the growth and differentiation of a number of different cells types including but not limited to fibroblasts, muscle cells, osteoblasts, keratinocytes and hepatocytes. In a preferred embodiment, the protein of the invention or part thereof is used in in vitro cultures such as those used for synthesis of recombinant proteins. The protein of the invention or part thereof is added to the culture in an amount effective to stimulate proliferation and/or differentiation. A more preferred application of this technique relates to the use of the protein of the invention or part thereof in generating or reparing in vitro tissues and organs such as but not limited to skin, cartilage,

and bone for transplants and grafts (Lancet – 1981, 1(8211):75-8), which disclosure is hereby incorporated by reference in its entirety).

In another embodiment, an antagonist of the protein of SEQ ID NO:296 may be administered to a subject to treat or prevent a cell proliferative disorder. Such disorders may include, but are not limited to, arteriosclerosis, atherosclerosis, bursitis, cirrhosis, hepatitis, mixed connective tissue disease (MCTD), myelofibrosis, paroxysmal nocturnal hemoglobinuria, polycythemia vera, psoriasis, primary thrombocythemia, and cancers including adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus. In one aspect, an antibody which specifically binds to the protein of the invention may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the protein of the invention. In another example, antisense nucleotides, triple helices, Genetic Suppressor Elements (GSE), ribozymes designed from nucleotides encoding the protein of the invention or part thereof using any methods to those skilled in the art are administrated to inhibit the expression of the protein of the invention.

Protein of SEQ ID NO: 410 (internal designation 179-9-4-0-B8-CS)

The protein of SEQ ID NO: 410 encoded by cDNA of SEQ ID NO: 169 found in fetal kidney is homologous to the proteins of ankyrin family protein and the proteins containing a characteristic ankyrin repeated motif (pfam accession number : PF00023). The protein of the invention shows homology with human ankyrin proteins (PIR accession number A35049 ; SP-TREMBL accession number : Q99407)), ankyrins from several different eucaryote species (*Drosophila melanogaster* : STR accession number Q9VAU5 ; mouse : STR accession number Q61302 and SwissProt accessiion number Q02357 ; cow : STR accession number AAF61702 ; *Arabidopsis thaliana* : STR accession number Q9ZQ79) and ankyrins from procaryote species (*Paramecium bursaria Chlorella* virus : STR accession number STR Q41164).

In addition, the protein of the invention shows homology with other proteins containing ankyrin repeat motif. The ankyrin repeat motif is a 33 amino acid motif and has an L-shaped structure consisting of two alpha helices following the beta hairpin loop (Gorina et al., Science.274-1005 (1996)). Examples of proteins comprising ankyrin repeats include: channels , enzymes toxins , transcription factors (Palek et al., Semin. Hematol. 27:290-332 (1990)), tankyrase (Smith et al., Science.282:1484-1487 (1998)) , multiple proteins involved in signal transduction, in particular integrin-linked kinases (Huang et al., Int. Mol. Med. 3:563-572 (1999)), inhibitors of cyclin-dependent kinases (Baumgartner et al., Structure. 6:1279-1290 (1998)), death-associated protein

kinase involved in apoptosis (Raveh et al., Proc. Nath. Acad. USA. 15:1572-1577 (2000)) and many others.

The ankyrin motif is also found in the protein of the invention (position 47 to 79).

Ankyrins are peripheral membrane proteins which have been found in erythrocyte, kidney and neuronal cells of mammals. Cells contain a cytoskeleton that links intracellular compartments with each other and the plasma membrane. Associations between the cytoskeleton and the lipid membranes bounding these compartments involve spectrin, ankyrin, and integral membrane proteins. Spectrin is a major component of the cytoskeleton and acts as a scaffolding protein. Similarly, ankyrin acts to tether the actin-spectrin moiety to membranes and to regulate the interaction between the cytoskeleton and membranous compartments. Different ankyrin isoforms are specific to different organelles and provide specificity for this interaction. Genes coding for three different mammalian ankyrins ($ankyrin_R$, $ankyrin_B$ and $ankyrin_G$) have been cloned. $Ankyrin_R$ was originally identified as part of the erythrocyte membrane skeleton, and was recently also localized to the plasma membrane of a subpopulation of post mitotic neurons in rat brain (Lambert, et al., 1993, J. Neurosci., 13, 3725-3735). $Ankyrin_B$ is a developmentally regulated human brain protein which has two alternatively spliced isoforms with molecular masses of 220 kilodaltons (kD) and 440 kD (Kunimoto, et al., 1991, J. Cell Biology, 115, 1319-1331). $Ankyrin_G$ is a more recently isolated human gene that encodes two neural-specific ankyrin variants (480 kD and 270 kD), which have been localized to the axonal initial segment and node of Ranvier (Kordeli, et al., 1995, J. Biol. Chem., 270, 2352-2359). Studies on mammalian ankyrins indicate that ankyrins bind a variety of proteins which have functions involved with the anion exchanger (Drenckhahn, et al., 1988, Science, 230, 1287-1289), Na+/K+-ATPase, amiloride-sensitive sodium channel in kidney (Smith, et al., 1991, Proc. Natl. Acad. Sci. U.S.A., 88, 6971-6975), voltage dependent sodium channel of the brain and the neuromuscularjunction (Srinivasan, et al., 1988, Nature, 333, 177-180), and nervous system cell adhesion molecules (Davis, et al., 1994, J. Biol. Chem., 269, 27163-27166).

Analyses of mammalian ankyrins have revealed that these large proteins are divided into three functional domains. These include an N-terminal membrane-binding domain of about 89-95 kD, a spectrin-binding domain of about 62 kD, and a C-terminal regulatory domain of about 50-55 kD. The membrane-binding domain is primarily comprised of tandem repeats of about 33 amino acids each. This domain usually has about 22-24 copies of these repeats. The repeat units appear to function in binding to membrane proteins such as anion exchangers, sodium channels, and certain adhesion molecules. The spectrin-binding domain, as the name implies, functions in binding to the spectrin-based cytoskeleton of cells positioned inside the plasma membrane. Finally, the regulatory domain, which is the most variant domain among the different ankyrins that have been studied, appears to function in as a repressor and/or an activator of the protein-binding activities of the other two domains. Some of the variability seen in this domain among different ankyrin species appears to be the result of alternative splicing of nascent transcripts. The regulatory domain can respond to

cellular signals, allowing remodeling of the cytoskeleton during the cell cycle and differentiation (Lambert, S. and Bennett, V. (1993) Eur. J. Biochem. 211:1-6). Ankyrin may be target for action of parasites. Erythrocyte ankyrin is cleaved by parasite proteases of plasmodium falciparum destabilizing erythrocyte membrane skeleton which facilitates parasite release (Raphael et al., Mol Biochem Parasitol. 110(2):259-272 (2000)). Recently, novel ankyrin proteins have been isolated from Dirofilaria and Brugia which may be useful in protecting animals, including humans, from diseases caused by parasitic helminths (United States Patent No. 6,063,599).

Ankyrin sequences have been identified in various libraries, at least 50% of which are associated with cancer and at least 23% of which are associated with the immune response. Of particular note is the expression of ANFP in reproductive and hematopoietic/immune, and gastrointestinal tissues. See United States Patent 5,989,863.

It is believed that the protein of SEQ. ID. NO: 410 is a member of the family of human ankyrin proteins and as such plays a role in regulating the interaction between the cytoskeleton and membranous components. The identification of a new member of the ankyrin family and the polynucleotides encoding it addresses a need in the art by providing new compositions which are useful in the diagnosis, prevention, and treatment of autoimmune/inflammatory, cell proliferative, vesicle trafficking disorders and in modulating the response to infectious diseases.

Preferred polypeptides of the invention are polypeptides comprsing the amino acids from positions 47 to 79. Other preferred polypeptides are fragments of SEQ.ID.NO: 410 having the desired biological activity. Further included in the invention are the polypeptides encoded by the human cDNA of clone 179-9-4-0-B8-CS. The polypeptides of SEQ ID NO: 410 may be interchanged with the corresponding polypeptides encoded by the human cDNA of clone 179-9-4-0-B8-CS. Further included in the invention are polynucleotides encoding said polypeptides. Preferred polynucleotides are those of SEQ ID NO: 169 and of the human cDNA of clone 179-9-4-0-B8-CS.

The invention also encompasses variants of the protein of the invention. A preferred variant is one which has at least about 80%, more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence of SEQ.ID.NO: 410, and which contains at least one functional or structural characteristic of ankyrin.

In a particular embodiment, the invention encompasses a polynucleotide sequence comprising the sequence of SEQ ID NO: 410, as well as variants of that sequence. Variants which encode at least one functional region characteristic of the ankyrin protein of the present invention are encompassed. Codon usage may be varied according to standard techniques in order to enhance expression in various hosts.

In one embodiment, the protein of SEQ.ID.NO: 410 or a fragment or derivative thereof may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of ankyrin. Examples of such disorders include, but are not limited to,

autoimmune/inflammatory disorders such as acquired immunodeficiency syndrome (AIDS), Addison's disease, adult respiratory distress syndrome, allergies, ankylosing spondylitis, amyloidosis, anemia, asthma, atherosclerosis. autoimmune hemolytic anemia, autoimmune thyroiditis, autoimmune polyenodocrinopathy-candidiasis-ectodermal dystrophy (APECED), bronchitis, cholecystitis, contact dermatitis, Crohn's disease, atopic dermatitis, dermatomyositis, diabetes mellitus, emphysema, episodic lymphopenia with lymphocytotoxins, erythroblastosis fetalis, erythema nodosum, atrophic gastritis, glomerulonephritis, Goodpasture's syndrome, gout, Graves' disease, Hashimoto's thyroiditis, hypereosinophilia, irritable bowel syndrome, multiple sclerosis, myasthenia gravis, myocardial or pericardial inflammation, osteoarthritis, osteoporosis, pancreatitis, polymyositis, psoriasis, Reiter's syndrome, rheumatoid arthritis, scleroderma, Sjogren's syndrome, systemic anaphylaxis, systemic lupus erythematosus, systemic sclerosis, thrombocytopenic purpura, ulcerative colitis, uveitis, Werner syndrome, complications of cancer, hemodialysis, and extracorporeal circulation, viral, bacterial, fungal, parasitic, protozoal, and helminthic infections, and trauma; cell proliferative disorders such as actinic keratosis, arteriosclerosis, bursitis, cirrhosis, hepatitis, mixed connective tissue disease (MCTD), myelofibrosis, paroxysmal nocturnal hemoglobinuria, polycythemia vera, psoriasis, primary thrombocythemia, and cancers including adenocarcinoma, leukemia, lymphoma, melanoma, myeloma, sarcoma, teratocarcinoma, and, in particular, cancers of the adrenal gland, bladder, bone, bone marrow, brain, breast, cervix, gall bladder, ganglia, gastrointestinal tract, heart, kidney, liver, lung, muscle, ovary, pancreas, parathyroid, penis, prostate, salivary glands, skin, spleen, testis, thymus, thyroid, and uterus; and vesicle trafficking disorders such as cystic fibrosis, glucose-galactose malabsorption syndrome, hypercholesterolemia, diabetes mellitus, diabetes insipidus, hyper- and hypoglycemia, Grave's disease, goiter, and Cushing's disease, ulcerative colitis, and gastric and duodenal ulcers.

In another embodiment, a vector capable of expressing the protein of SEQ.ID.NO: 410 or a fragment or derivative thereof may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of ankyrin including, but not limited to, those described above.

In a further embodiment, a pharmaceutical composition comprising a substantially purified protein of SEQ.ID. NO. 410 or a portion of the protein in conjunction with a suitable pharmaceutical carrier may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of the same or a similar protein including, but not limited to, those provided above.

In still another embodiment, an agonist of the protein of SEQ. ID. NO. 410 which modulates the activity of the protein may be administered to a subject to treat or prevent a disorder associated with decreased expression or activity of the protein, or other ankyrin proteins, including, but not limited to, those listed above.

In another embodiment, the polypeptide of SEQ. ID. NO. 410 may be used to produce antagonists using methods which are generally known in the art. In particular, purified polypeptide may be used to produce antibodies or to screen libraries of pharmaceutical agents to identify those which specifically bind ankyrin proteins. Neutralizing antibodies (i.e., those which inhibit dimer formation) can also be prepared for therapeutic use.

In a further embodiment, an antagonist of the protein of SEQ. ID. NO. 410 may be administered to a subject to treat or prevent a disorder associated with increased expression or activity of the same protein or other members of the ankyrin family of proteins. Such disorders may include, but are not limited to, those discussed above. In one aspect, an antibody which specifically binds the claimed polypeptide may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the polypeptide.

In an additional embodiment, a vector expressing the complement of the polynucleotide of SEQ. ID. NO. 169 may be administered to a subject to treat or prevent a disorder associated with increased expression or activity of ankyrin proteins including, but not limited to, those described above.

In other embodiments, any of the proteins, antagonists, antibodies, agonists, complementary sequences, or vectors of the invention may be administered in combination with other appropriate therapeutic agents. The combination of therapeutic agents may act synergistically to effect the treatment or prevention of the various disorders described above. Using this approach, one may be able to achieve therapeutic efficacy with lower dosages of each agent, thus reducing the potential for adverse side effects.

In another embodiment of the invention, the polynucleotides encoding the polypeptide of SEQ. ID. NO. 410, or any fragment or complement thereof, may be used for therapeutic purposes. In one aspect, the complement of the polynucleotide encoding the above-identified polypeptide may be used in situations in which it would be desirable to block the transcription of the mRNA. In particular, cells may be transformed with sequences complementary to polynucleotides encoding the polypeptide. Thus, complementary molecules or fragments may be used to modulate the activity of the claimed polypeptide or related ankyrin proteins, or to achieve regulation of gene function.

In another embodiment of the invention, the nucleotide sequence encoding the polypeptide of SEQ. ID. NO. 410 can be used to turn off the genes expressing the polynucleotide or related ankyrin proteins. In particular, a cell or tissue can be transformed with expression vectors which express high levels of the polynucleotide, or fragment thereof. Such constructs may be used to introduce untranslatable sense or antisense sequences into the cell. Expression vectors derived from retroviruses, adenoviruses, or herpes or vaccinia viruses, or from various bacterial plasmids, may be used for delivery of the nucleotide sequences to a targeted organ, tissue, or cell population.

An additional embodiment of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic effects discussed above. The composition may be delivered via a variety of different routes.

In another embodiment, antibodies which bind the polypeptide of SEQ. ID. NO. 410 may be used for the diagnosis of disorders characterized by expression of ANFP, or in assays to monitor patients being treated with the polypeptide, other ankyrin proteins or agonists, antagonists, or inhibitors of the same. A variety of assay types, including ELISAs, RIAs, and FACS, can be used.

In another embodiment of the invention, the polynucleotide of SEQ. ID. NO. 169 itself, may be used for diagnostic purposes. The polynucleotide can be used to generate oligonucleotide sequences, complementary RNA and DNA molecules, and PNAs which are useful in diagnosis. The polynucleotide and related molecules may be used to detect and quantitate gene expression in biopsied tissues in which expression of the polypeptide of SEQ. ID. NO. 410 or other ankyrin polypeptides may be correlated with disease. The diagnostic assay may be used to determine absence, presence, and excess expression of the polypeptides, and to monitor regulation of polypeptide levels during therapeutic intervention. Examples of diagnostic methods include Southern or Northern analysis, dot blot, or other membrane-based technologies; in PCR technologies; in dipstick, pin, and multiformat ELISA-like assays; and in microarrays utilizing fluids or tissues from patients to detect altered ANFP expression. Such qualitative or quantitative methods are well known in the art. Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual patient.

In further embodiments, oligonucleotides or longer fragments derived from any of the polynucleotide sequences described herein may be used as targets in a microarray. The microarray can be used to monitor the expression level of large numbers of genes simultaneously and to identify genetic variants, mutations, and polymorphisms. This information may be used to determine gene function, to understand the genetic basis of a disorder, to diagnose a disorder, and to develop and monitor the activities of therapeutic agents.

In another aspect of the invention, the polypeptide may be used to stimulate the expression of genes that have a role in organ and organ system development. Thus, in a preferred embodiment, the protein of the invention, a fragment, or derivative thereof, may be administered to a subject to treat or prevent developmental disorders.

In a further embodiment, the protein of the invention may be administered to a subject to treat or prevent a cardiovascular disorder. Such disorders can include, but are not limited to, arteriosclerosis including atherosclerosis and nonatheromatus arteriosclerosis, hypertension, stroke, coronary artery disease, ischemia, myocardial infarction, angina pectoris, cardiac arrhythmias, sinoatrial node blocks, atrioventricular node blocks, chronic hemodynamic overload, aneurysm,

Jervell and Lange-Nielsen syndrome, and long QT syndrome. The protein of the invention may also be used as a marquer of cardiac hypertrophy so it may be included in diagnosis kit for this disease.

In another embodiment of the invention, the polypeptide and/or polynucleotide may be used to inhibit cellular proliferation and to treat and/or diagnose disorders associated with cellular proliferation including but not limited to cancer.

In a further embodiment of the invention, an antagonist of the protein of the invention may be administered to a subject to treat or prevent a cancer. In one aspect, an antibody which specifically binds the protein of the invention may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the protein of the invention.

In yet another embodiment, an antagonist of the protein of the invention may be administered to a subject to treat or prevent a neuronal disorder. Such a disorder may include, but is not limited to, akathesia, Alzheimer's disease, amnesia, amyotrophic lateral sclerosis, bipolar disorder, catatonia, cerebral neoplasms, dementia, depression, diabetic neuropathy, Down's syndrome, tardive dyskinesia, dystonias, epilepsy, Huntington's disease, peripheral neuropathy, multiple sclerosis, neurofibromatosis, Parkinson's disease, paranoid psychoses, postherpetic neuralgia, schizophrenia, and Tourette's disorder. In one aspect, an antibody which specifically binds the protein of the invention may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the protein of the invention.

In another embodiment, the protein of the invention can be administered to a subject to treat or prevent a malaria. The protein of the invention may be used also in diagnosis of malaria.

In yet another embodiment, the polynucleotide and/or the polypeptide of the present invention can be used as a therapeutic composition capable of protecting an animal from a disease caused by a parasitic helminth. The polypeptide can be used a target for antiparasitic vaccines and drugs.

Ankyrin has been shown to underlie membrane proteins including CD44, the voltage-dependent sodium channel, NA+/K+ ATPase and the anion exchanger protein. It is believed that the formation of a direct connection between ankyrin and functionally important transmembrane proteins/membrane skeleton may be one of the earliest events to occur during signal transduction and cell activation. Thus, in a further embodiment, the polypeptide of the present invention can be used to disrupt the connection between ankyrin and the membrane thus affecting fundamental processes within the cell.

The polypeptide of the present invention can be further used to screen for compounds that inhibit or enhance the binding of ankyrin binding proteins and to affect the association between ankyrin and proteins, such as Alpha-Na,K-ATPase, which are critical to intracellular transport of ions and nutrients.

In yet another embodiment the regulatory domain of the polypeptide of SEQ. ID.NO. 410 or antagonists thereof can be used to enhance or disrupt the protein binding activities of the other domains.

Proteins of SEQ ID NOs: 385 and 416 (internal designations 105-021-3-0-C3-CS and 188-31-1-0-E6-CS respectively)

The 354 amino acids protein of SEQ ID NO: 385 encoded by the cDNA of SEQ ID NO: 144 found in brain displays 6 kelch motifs (pfam accession number PF01344) at positions 20-66, 68-114, 116-162, 164-209, 211-265 and 270-316. Morevoer, 4 residues conserved in over 90% of kelch family sequences are found in the protein of invention: di-glycine at positions 133-134, tyrosine 148 and tryptophan 155. In addition, six residues separate the tyrosine 148 and the tryptophan 155: this feature is conserved in over 70% of kelch proteins (Adams et al., trends in cell biology, 10:17-24, 2000). The proteins of the invention encoded by the cDNA of SEQ ID NO: 144 is a polymorphic variant of the protein of SEQ ID NO: 416 encoded by the cDNA of SEQ ID NO: 175, thought to have the same functions and utilities.

Drosophila kelch is located in the ring canals which are actin-rich bridges. Kelch localizes to the rim of preformed canals and serves to maintain actin organization (Xue et al., Cell (72)681-693, 1993; Robinson et al., J. Cell Biol. (138)799-810, 1997). In mammalian sperm, calicin is located within an actin-negative structure termed the calyx, which is involved in the morphogenesis of the spermatocyte (von Bulow et al., Exp. Cell Res. (219)407-413, 1995). Calcin and a well-structured calyx are both lacking in certain teratozoospermias, possibly indicating a central role for calicin in the organization of this structre (Courtot et al., Mol. Reprod. Dev. (28)272-279, 1991). In *Schizosaccharomyces pombe*, Ral2p acts down-stream of Ras1p in pathways that affect cell morphology, conjugation and sporulation. The spherical morphology and mating defects of ral2-null cells are complemented by overxpression of Ras1p, indicating a close functional interaction between the two proteins (Fukui et al., Mol. Cell. Biol. (9)5617-5622, 1989). The transcription factor Nrf2 is sequestered by the kelch-repeat containing Keap1 protein under normal cellular conditions. The stimulation by agents such as diethylmaleate induces the translocation of Nrf2 to the nucleus to initiate the cytoprotective electrophilic counterattack response (Itoh et al., Genes Dev. (13)76-86, 1999). Lytic infection of cells by herpes simplex virus is initiated by binding of virally encoded VP16 to HCF-1, a protein thought to have a normal role in cell-cycle progression. The HCF-VP16 complex then assembles with Oct-1 transcription factor on cis-regulatory targets in the HSV genome to initiate virus replication. (Wilson et al., Mol. Cell. Biol (17)6139-6146, 1997; Hughes et al., J. Biol. Chem. (274)16437-16443, 1999). Two recently discovered mammalian kelch-repeat proteins have extracellular roles. Human attractin appears to participate in normal immune defence as a serum glycoprotein released by activated T cells. In coculture assays, attractin stimulates adhesion and spreading of monocytes, facilitating the development of T-cell clusters and

cellular immune responses (Duke-Cohan et al., Proc. Natl. Acad. Sci. U. S. A. (95)11336-11341, 1998). Attractin is orthologous to the extracellular domain of mouse mahogany, a large, multidomain, transmembrane protein that has been implicated in the homeostasis of energy metabolism by its suppressive effects on certain types of obesity in mice (Gunn et al., Nature (398)152-156, 1999; Nagle et al., Nature (398)148-152, 1999).

Evidence for the importance of kelch repeat beta-propellers in protein function has also come from studies of natural an engineered loss-of-function mutations. *Caenorhabditis elegans* Spe-26 mutant spermatocytes fail to complete meiosis, contain multiple nuclei and show gross disorganization of actin filaments and organelles. For five out of the six alleles that have been examined in detail, the mutations map within the kelch repeats (Varkey et al., Genes Dev. (9)1074-1086, 1995). Of particular interest are the point mutations in RAG-2 that have been identified in some cases of human B-cell-negative severe combined immuno-deficiency or of Omenn syndrome (Schwarz et al., Science (274)97-99, 1996; Villa et al., Cell (93)885-896, 1998). Very recently, the gigaxonin, a new member of the cytoskeletal BTB/kelch repeat family, is described as mutated in giant axonal neuropathy (Bomont et al., Nat. Genet. (26) 370-374, 2000).

It is believed that the proteins of the invention are members of the kelch superfamily and, such as, play a role in the association with the actin cytoskeleton, the organization of cytoskeletal, plasma membrane or organelle structures, the coordination of morphology and growth, the gene expression, the viral pathogenesis, the immune defence. In particular, the proteins of invention are highly expressed in brain and is believed to be related to the CNS disorders. Preferred polypeptides of the invention are polypetides comprising the amino acids of the proteins of invention at positions 20-66, 68-114, 116-162, 164-209, 211-265 and 270-316. In one embodiment, the proteins of the invention or part thereof are used to modulate actin organization in cells thus affecting the cytoskeleton and cell function in general.

The invention also features compounds, e.g., proteins, which interact with the protein of the invention. Any method suitable for detecting protein-protein interactions may be employed for identifying transmembrane proteins, intracellular, or extracellular proteins that interact with the protein. Among the traditional methods which may be employed are co-immunoprecipitation, crosslinking and co-purification through gradients or chromatographic columns of cell lysates, or proteins obtained from cell lysates, and the use of the proteins of the invention to identify proteins in the lysate that interact with it. For these assays, the protein of the invention can be full length or some other suitable protein polypeptide fragment. Once isolated, such an interacting protein can be identified and cloned and then used, in conjunction with standard techniques, to identify proteins with which it interacts. For example, at least a portion of the amino acid sequence of a protein which interacts with the protein of the invention can be ascertained using techniques well known to those of skill in the art, such as via the Edman degradation technique. The amino acid sequence obtained may be used as a guide for the generation of oligonucleotide mixtures that can be used to

screen for gene sequences encoding the interacting protein. Screening may be accomplished, for example, by standard hybridization or PCR techniques. Techniques for the generation of oligonucleotide mixtures and the screening are well-known. ("PCR Protocols: A Guide to Methods and Applications," Innis et al., eds. Academic Press, Inc., NY, 1990).

Additionally, methods can be employed which result directly in the identification of genes which encode proteins that interact with the protein of the invention. These methods include, for example, screening expression libraries, in a manner similar to the well known technique of antibody probing of lambda.gt11 libraries, using labeled polypeptide or a protein fusion protein, e.g., a protein polypeptide or domain fused to a marker such as an enzyme, fluorescent dye, a luminescent protein, or to an IgFc domain.

Another embodiment of the invention relates to compositions and methods using the protein of the invention or part thereof to modulate actin organization and related cytoskeletal protein organization in cells and in particular, cells of the CNS. Compositions containing the protein of the invention and fragments thereof may be therapeutic and used to treat a variety of neuronal disorders. An additional embodiment of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic effects discussed. The composition may be delivered via a variety of different routes.

In yet another embodiment, an antagonist of the protein of the invention may be administered to a subject to treat or prevent a neuronal disorder. Such a disorder may include, but is not limited to, akathesia, Alzheimer's disease, amnesia, amyotrophic lateral sclerosis, bipolar disorder, catatonia, cerebral neoplasms, dementia, depression, diabetic neuropathy, Down's syndrome, tardive dyskinesia, dystonias, epilepsy, Huntington's disease, peripheral neuropathy, multiple sclerosis, neurofibromatosis, Parkinson's disease, paranoid psychoses, postherpetic neuralgia, schizophrenia, and Tourette's disorder. In one aspect, an antibody which specifically binds the protein of the invention may be used directly as an antagonist or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissue which express the protein of the invention.

Another embodiment of the invention encompasses DNA sequences which encode the proteins of the invention that may be derived through biological or synthetic chemistry processes. Polynucleotides sequences capable of hybridizing to the cDNA sequences of SEQ ID NOs: 144 and 175are also included in the scope of the invention.

Further included in the invention are the polypeptides encoded by the human cDNA of clones 105-021-3-0-C3-CS and 188-31-1-0-E6-CS. The polypeptides of SEQ ID NOs: 385 and 416 may be interchanged with the corresponding polypeptides encoded by the human cDNA of clones 105-021-3-0-C3-CS and 188-31-1-0-E6-CS. Further included in the invention are polynucleotides

encoding said polypeptides. Preferred polynucleotides are those of SEQ ID NOs: 144 and 175 and of the human cDNA of clones 105-021-3-0-C3-CS and 188-31-1-0-E6-CS.

Another embodiment of the invention to methods of using the nucleotidic sequence or part thereof of invention to search homologs to the protein of invention. The sequence can be used as template of PCR reactions, allowing the detection/quantification of the protein of invention or part of thereof or the homologs. The complementary sequence or part of thereof may be used as hybridization probes to detect/quantify the transcription level, as well in *in vitro* level as the cellular level. In particular, such probes may be used in a diagnostic context, for example in the cellular or tissue *in situ* hybridization.

Another embodiment of the invention to methods of using the nucleotidic sequence or part thereof of invention to design antisense oligonucleotides to modulate the *in vitro* or *in vivo* expression of the protein or the part or thereof gene expression. This may be useful in therapeutic area of diseases listed above, particularly in the context where the protein of invention is expressed in abnormally high level.

In a further embodiment of the invention, the protein of invention or portions thereof are used to produce specific antibodies. These antibodies may have applications in the diagnostics, the purification of the protein of invention or part of thereof or a homolog. They may also help to vizualize the locations of the proteins associated to the protein of invention in the cell, in particular for highlighting structure-related proteins. The methods described herein in which protein antibodies are employed may be performed, for example, by utilizing pre-packaged diagnostic kits comprising at least one specific cDNA of SEQ ID NO. or antibody reagent described herein, which may be conveniently used, for example, in clinical settings, to diagnose patients exhibiting symptoms of various CNS disorders.

In still another preferred embodiment, the present inventions relates to methods of using the protein of the invention or part thereof in gene therapy, particularly in the diseases involving the CNS, particularly in the context where the protein of the invention is expressed in abnormally low level. Gene therapy is a potential therapeutic approach in which normal copies of the cDNAs of SEQ ID NOs: 144 and 175 may be introduced into subjects to successfully code for normal protein in several different affected cell types.

Another aspect of the present invention includes a formulation comprising the protein of the invention or part thereof and a pharmaceutically or physiologically acceptable carrier. A formulation of the present invention comprises a combination of one or more peptides as described herein, or mimetopes thereof; a combination of antibodies as described herein, or mimetopes thereof; or a combination of antibodies and peptides as described herein, or mimetopes thereof. Such a formulation may be administered to a subject in need thereof to treat or prevent a disorder associated with decreased expression or activity of the protein. Examples of such disorders include

but are not limited to those of the CNS and other tissues where the association of actin appears to be abnormal.

Proteins of SEQ ID NO: 391, 393, 405 and 407 (internal designations 145-52-2-0-D12-CS, 145-7-3-0-D3-CS, 174-17-1-0-D6-CS and 174-38-4-0-D11-CS respectively)

The cluster of four proteins (SEQ ID NOs: 391, 393, 405 and 407) encoded by the cDNAs of SEQ ID NOs: 150, 152, 164 and 166 respectively exhibit very strong homology to claudin-8, a member of PMP22-Claudin family (PF00822). SEQ ID NO: 405 (174-17-1-0-D6-CS) shares a high degree of identity with human claudin-8. SEQ ID NO: 393 contains two amino acid substitutions as compared to human claudin-8 (T129A and S151P); thus, it appears to be a polymorphic variant of claudin-8. SEQ ID Nos: 393 and 405 contain four membrane spanning segments.

SEQ ID NOs: 391 and 407 are polymorphic forms of claudin-8. The protein of SEQ ID NO: 391 (145-52-2-0-D12-CS) is 162 amino-acids long, contains three theoretical membrane spanning segments, and shows three amino acid substitutions as compared to the previously identified claudin-8 protein (R31I, S151P and E162). The protein of SEQ ID NO: 407 (174-38-4-0-D11-CS) is 43 amino-acids long. SEQ ID NO: 407 contains a stop codon at position 44 and contains no apparent membrane spanning segments.

The Claudin family of proteins comprises more than twenty (20) small glycoproteins with four predicted transmembrane domains. The tissue distribution pattern of claudins varies significantly, depending on claudin species. Many have been identified as components of tight junction (TJ) strands which contribute in regulation of cell polarity and permeability. Polarized epithelial and endothelial cells form barriers that separate biological compartments and regulate homeostasis. The tight junction (TJ) is a specialized membrane domain at the most apical region of polarized epithelial and endothelial cells that not only creates a primary barrier to prevent paracellular transport of solutes (barrier function) but also restricts the lateral diffusion of membrane lipids and proteins to maintain the cellular polarity (fence function). Tight junctions appear to represent a continuous network of interconnected rows of intramembranous particles that appear as strands with complementary grooves. The TJ-specific integral membrane proteins, i.e. the components of TJ strands, occludins and claudins, were only recently identified.

Claudin-1 and -2 have the ability to induce the formation of networks of strands/grooves at cell-cell contact sites when introduced into fibroblasts lacking TJs. Occludin induces only a small number of short strands at cell-cell contact sites in fibroblasts, thus, it is an accessory protein in terms of TJ strand formation. Claudin transfection experiments in fibroblasts revealed the TJ strand itself can be formed without occludin (Saitou M et al. J. Cell Biol. 141: 397-408 (1998), Furuse M et al. J. Cell Biol. 143: 391-401 (1998)).

Initially several members of the claudin family were reported (RVP1, Clostridium perfringens enterotoxin receptor (CPE-R), and TMVCF (transmembrane protein deleted in Velo-cardio-facial

syndrome)), but their physiological functions were not determined. After the identification of claudin-1 and -2 as novel components of TJ strands (Furuse, M. et al. J. Cell Biol. 141, 1539-1550 (1998)), CPE-R was shown to remove specific claudins from TJs. In its presence, TJ strands in C3L cells gradually disintegrate and the number of TJ strands and the complexity of their network decreases markedly (Sonoda N et al. J Cell Biol 147(1):195-204 (1999)). In distal tubules of the kidney, claudin-4 (CPE-R) and claudin-8 were co-localized with occludin at their junctional complex region. In liver, claudin-3 and occludin were co-localized along bile canaliculi and TJ strands were labeled heavily and specifically with anti-claudin-3 Ab (Morita et al. PNAS 96 (2): 511-516 (1999)). The claudins have been shown to create the paracellular diffusion barrier and, surprisingly, they may also confer channel-like selectivity for passage of solutes through the tissue barrier (Anderson JM and Christina M. Van Itallie CM, Current Biology 9:R922-R924 (1999)).

The existence of the claudin multigene family as well as the tissue distribution pattern of each claudin species suggests that similar complexity can be expected in TJs and contributes to the generation of functional diversity of TJs in vivo. More than two distinct claudins are co-expressed in single epithelial cell. Claudins interact between each of the paired strands in a heterophilic manner and distinct claudins are (except in some combinations) co-incorporated into individual TJ strands (Furuse M et al. J Cell Biol 147(4):891-903 (1999)).

Several claudins have been shown to be expression markers of malignant cells. For example, SEMP1 (senescence-associated epithelial membrane protein) is expressed in normal human tissues, including adult and fetal liver, pancreas, placenta, adrenals, prostate and ovary; however, SEMP1 is expressed at low or undetectable levels in a number of human breast cancer cell lines (Swisshelm K et al. Gene 226:285-295 (1999)). Another member of the claudin family was found to be exclusively expressed in MCF-7ADR mammary carcinoma cells. MCF-7ADR carcinoma cells are estradiol-independent for growth, estrogen-receptor negative, tamoxifen resistant, vimentin positive and invasive in vitro and in vivo (Schiemann S et al., Anticancer Res 17(1A):13-20 (1997)). Further, down regulation of the expression of claudin-1 has been associated with oncogenesis in rat salivary gland epithelium cells (Li D and Mrsny RJ J Cell Biol 148(4):791-800 (2000)).

The increase in microvascular permeability in human gliomas, contributing to clinically severe symptoms of brain edema, appears to be the result of a dysregulation of junctional proteins. Increased TJ permeability of the colon epithelium, and consequently a decrease in epithelial barrier function, precedes the development of colon tumors, including carcinomas and adenomatous polyps (Soler AP et al. Carcinogenesis 20(8):1425-1431 (1999)). Studies of the interendothelial junctions in tumor microvessels of human glioblastoma multiforme show that the expression of claudin-1 is lost in the majority of tumor microvessels, whereas claudin-5 is significantly down-regulated only in hyperplastic vessels. A relationship between claudin-1 suppression and the alteration of tight junction morphology is likely to correlate with the increase of endothelial permeability (Liebner S et al. Acta Neuropathol (Berl) 100(3):323-331 (2000)).

The human phenotype of mutations in claudin-16 suggests that it creates a channel allowing magnesium to diffuse through renal tight junctions. Similarly, a mouse knockout of claudin-11 reveals its role in formation of tight junctions in myelin and between Sertoli cells in testis (Mitic LL et al. Am J Physiol Gastrointest Liver Physiol 279(2):G250-254 (2000)).

Opening of TJs by environmental proteinases may be the initial step in the development of asthma to a variety of allergens. The lung epithelium forms a barrier that allergens must cross before they can cause sensitization. The cysteine proteinase allergen Der p 1 from fecal pellets of *Dermatophagoides pteronyssinus* (the house dust mite (HDM)) causes disruption of intercellular tight junctions (TJs), which are the principal components of the epithelial paracellular permeability barrier. TJ breakdown nonspecifically increases epithelial permeability, allowing Der p 1 to cross the epithelial barrier. Putative Der p 1 cleavage sites were found in peptides from an extracellular domain of claudin-1. House dust mite (HDM) allergens are important factors in the increasing prevalence of asthma (Wan H et al. J Clin Invest 104(1):123-33 (1999)).

In many intestinal and systemic diseases, intestinal barrier damage is marked by changes in intestinal permeability which are, in turn, related to alteration in tight junction function (Gasbarrini G, Montalto MItal J Gastroenterol Hepatol 31(6):481-488 (1999)). Permeability of the tight junctions can be modified by bacterial toxins, cytokines, hormones and drugs. Oligodendrocyte-specific protein (OSP/claudin-11), found in CNS myelin, appears to be a promising candidate for auto-antigenic involvement in autoimmune demyelinating disease. The presence of anti-OSP Abs in the cerebrospinal fluid was reported for relapsing-remitting multiple sclerosis (MS). Murine OSP peptides elicit clinical experimental autoimmune encephalomyelitis in animal models for MS and induces mononuclear cell infiltrates and focal demyelination. Also OSP peptides elicit robust proliferative responses in T cells (Stevens DB et al. J Immunol 162:7501-7509 (1999)). OSP/claudin-11 appears to modulate proliferation and migration of oligodendrocytes, presumably through the membrane interactions at tight junctions and with the extracellular matrix (Bronstein JM et al. J Neurosci Res 59(6):706-711 (2000)). Recently claudin-11 has been shown to play a key role in the formation of hematotesticular barrier; it is regulated by FS hormone and by cytokines in early fetal and postnatal development in Sertoli cells (Hellani A. et al. Endocrinology 141: 3012-3019 (2000)).

SEQ ID NOs: 391, 393, 405 and 407 are new human proteins having biological activities described for claudins. Nucleic acids encoding the proteins of interest are over represented in fetal kidney and in salivary gland. The subject invention provides polynucleotides encoding the proteins of SEQ ID Nos: 391, 393, 405 and 407. In one embodiment, the polypeptides of SEQ ID NOs: 391, 393, 405 and 407 are interchanged by the polypeptides encoded by clones 145-52-2-0-D12-CS, 145-7-3-0-D3-CS, 174-17-1-0-D6-CS and 174-38-4-0-D11-CS. Also provided are use of these proteins, fragments, derivatives thereof (and related polynucleotides) for the diagnosis, treatment, or prevention of tumors and another diseases, including disorders associated with altered epithelial function. The invention also encompasses possible variants of the proteins of interest which have at least about 80%,

more preferably at least about 90%, and most preferably at least about 95% amino acid sequence identity to the amino acid sequence, provide the variants have at least one of the functional or structural characteristics of the identified claudin-like proteins.

In one embodiment of the subject invention, the proteins of interest, or biologically active fragments or variants thereof, may be administered to a subject to treat or prevent disorders of salivary gland, kidney and prostate. The subject invention also provides therapeutic regimens for the treatment of epithelial dysfunction and cancer.

The disorders which may be treated in accordance with the subject invention include, but are not limited to, asthma, eczema, atopic dermatitis, contact dermatitis, stasis dermatitis, seborrheic dermatitis, psoriasis, lichen planus, pityriasis rosea, acne vulgaris, acne rosacea, pemphigus vulgaris, pemphigus foliaceus, paraneoplastic pemphigus, bullous pemphigoid, herpes gestationis, dermatitis herpetiformis, linear IgA disease, epidermolysis bullosa acquisita, dermatomyositis, lupus erythematosus, scleroderma, and morphea; gastritis, peptic ulcers, cholelithiasis, cholecystitis, pancreatitis, cirrhosis, ulcerative colitis, Crohn's disease, and irritable bowel syndrome; Addison's disease, Lowe's syndrome, glomerulonephritis, chronic glomerulonephritis, tubulointerstitial nephritis, inherited X-linked nephrogenic diabetes insipidus, autosomal dominant polycystic kidney disease, autoimmune demyelinating disease, multiple sclerosis, glioma, and other tumors.

A further aspect of the invention provides a method for treating these and/or other pathological states by administering, to a patient, a therapeutically effective amount of one or more of the proteins of interest. The proteins of interest may, optionally, be simultaneously or sequentially administered in conjunction with cytokines and/or interleukins which have been shown to improve claudin expression.

In another embodiment, a vector capable of driving expression of one or more of the proteins of interest, or a biologically active fragment or variant thereof, may be administered to a subject to treat or prevent an epithelial permeability disorder including, but not limited to aforementioned disorders.

Another embodiment of the subject invention provides compositions and methods of treating, or reducing the incidence of, asthma comprising the administration of therapeutically effective amounts of the proteins of the subject invention. In one embodiment, purified fragments, or synthetically modified peptides, derived from the extracellular domains of the proteins of interest may be administered in the therapeutic regimen. The peptides, containing the putative cleavage sites for environmental allergen proteinases, may be administrated in amounts which competitively inhibit the proteinase activity of the allergen. The peptides may be designed to bind allergen, optionally in an irreversible manner, and inhibit proteinase activity.

The negative effects of the usual preservation solutions on epithelial and endothelial permeability in organs to be transplanted are generally known (Trocha S.D. et al. Ann.Surg. 230: 105-113 (1999)). Increases in permeability leads to tissue injury and edema. Disorganization of tight junctional proteins appears to be responsible for the observed tissue injury and edema. Thus, in another embodiment, purified proteins of interest, or variants and/or biologically active fragments thereof, may

be added in organ preservation solutions to maintain the content and integrity of tight junctions in organs.

In another embodiment, the subject invention provides methods of producing "bioartificial" epithelia from non-epithelial cells. The "bioartificial" epithelia produced according to the invention may be used for reconstructive surgical procedures, for treating of disorders related to epithelial loss (for hereditary, traumatic or oncological reasons) or for another therapeutic purposes (e.g., burn treatments). "Bioartificial" epithelial cells can be obtained by transfection and remodeling of the autologous patient cells not affected by any of the aforementioned disorders. The use of autologous cells in the preparation of the "bioartificial" epithelial cells of the invention in methods of treating disorders, conditions, or diseases associated with the loss of epithelial cells reduces or eliminates the risk of tissue rejection typically observed in transplantation methodologies. Methods of bioartificial tissue engineering are generally known to those skilled in the art (for a review, see Machens H.G. et al. Cells Tissues Organs 167: 88-94 (2000)).

In another embodiment of the subject invention provides antibodies which specifically bind to the proteins of SEQ ID Nos: 391, 393, 405 and 407. The antibodies may also specifically bind to fragments or variants of the proteins described in SEQ ID Nos: 391, 393, 405 and 407. The antibodies of the invention may be used to detect the protein of interest in human body fluids, extracts of cells or tissue extracts. The detection assays may be used for epithelial cancer prognosis and for the diagnosis of disorders. The assays may also be used to monitor patients being treated with the proteins of interest.

In another embodiment, the polynucleotide sequences, or fragments of said polynucleotide sequences, encoding the proteins of interest may be used for the identification or diagnosis of a disorder associated with expression of the proteins of SEQ ID Nos: 391, 393, 405 and 407. Hybridization assays which allow for the detection of polynucleotide sequences of the invention are well known to the skilled artisan. These assays include, and are not limited to, Northern blots, Southern blots, and PCR methodologies.

Another embodiment of the invention provides the proteins of SEQ ID Nos: 391, 393, 405 and 407, variants, immunogenic fragments, or biologically active fragments of said proteins for screening libraries of compounds in any of a variety of drug screening techniques. The proteins of SEQ ID Nos: 391, 393, 405 and 407, variants, immunogenic fragments, or biologically active fragments of said proteins employed in such screening may be free in solution, affixed to a solid support, recombinantly expressed on, or chemically attached to, a cell surface, or located intracellularly. The formation of binding complexes between the protein of interest and the agent being tested may be measured by methods well known to those skilled in the art.

Yet another embodiment of the invention provides methods of screening compounds which modulate epithelial permeability Polynucleotides encoding the proteins of SEQ ID Nos: 391, 393, 405 and 407, variants, immunogenic fragments, or biologically active fragments of said proteins, may be

recombinantly expressed in cells typically lacking TJs according to methods discussed supra. These cells may then be used to assess therapeutic modulators (based, for example, on CPE –like compounds) for the ability to increase or decrease epithelial cell permeability. Compounds identified in these modulator screen assays may then be used in therapeutic protocols to adjust epithelial cell permeability as desired by the practitioner.

The intestinal epithelium is a major barrier to the absorption of hydrophilic drugs. The presence of intercellular junctional complexes, particularly the tight junctions, renders the epithelium impervious to hydrophilic drugs, which cannot diffuse across the cells through the lipid bilayer of the cell membranes (Ward PD et al. Pharmaceutical Science and Technology Today 3:10:346-358 (2000)). Therefore, in another embodiment of the subject invention the proteins of SEQ ID Nos: A, B, C, or D, variants, or biologically active fragments of said proteins and their molecular partners can be used for the rational design of compounds that can effectively and safely increase paracellular permeability for selected drugs. For example, polynucleotides encoding the proteins of interest or any fragment or derivatives thereof, may be used for these purposes. In one aspect, the complement of the polynucleotide encoding the protein of interest may be used in situations in which it would be desirable to block the transcription of the mRNA encoding the proteins of interest, especially for temporally increasing epithelial permeability (useful for drug delivery). Alternatively, sense or antisense oligonucleotides may be designed from various locations along the coding or control regions of polynucleotide sequences encoding the proteins of SEQ ID Nos: 391, 393, 405 and 407, as well as variants, or biologically active fragments of said proteins to control expression of the proteins. Methods of producing and using sense and antisense oligonucleotides are well known to those skilled in the art.

Claudins are unique proteins with specific protein-binding properties. Therefore, in another preferred embodiment, the proteins of SEQ ID Nos: 391, 393, 405 and 407, variants, or biologically active fragments of said proteins may be used as a component of drug delivery vehicles such as colloids or liposomes. The proteins of the proteins of SEQ ID Nos: 391, 393, 405 and 407, variants, or biologically active fragments of said proteins may be incorporated into the lipid membranes of liposomes and can serve as specific targeting agents which bind the specific epithelial targets and facilitate targeted epithelium drug delivery. The methods of design of such type of drug delivery systems is known by those skilled in the art (Smith H.J. Introduction to the principles of drug design and action, 3rd ed. (1998)). Alternatively, active agents, such as chemotherapeutic agents, radioisotopes, prodrugs, may be directly attached, recombinantly or chemically, to the proteins of SEQ ID NOs: 391, 393, 405 and 407, variants, or biologically active fragments of said proteins and used in therapeutic regimens.

Proteins of SEQ ID Nos: 278, 282 and 300 (internal designations 160-37-2-0-H7-CS, 174-33-3-0-F6-CS, 184-4-1-0-A11-CS respectively)

The protein of SEQ ID No: 278 (and the corresponding allelic variants 282 and 300) encoded by the cDNA SEQ ID No: 37 (41, 59 respectively) shows homology to a human transmembrane protein (HTMN-23, Genseq accession number Y57899). The protein of SEQ ID No: 278 (and the corresponding polymorphic variants 282 and 300) overexpressed in salivary gland contains 9 potential transmembrane segments from positions 85 to 105, 116 to 136, 164 to 184, 187 to 207, 332 to 352, 376 to 396, 404 to 424, 465 to 485 and 499 to 519, thus displaying chracteristic features of type III transmembrane proteins (Singer, Annu. Rev. Cell Biol., 6:247-296, 1990). Furthermore, a predicted localisation in the endoplasmic reticulum (ER) is found for the protein of the invention with the software psort.

The normal fonctioning of the eukaryotic cell requires that all the newly synthesized proteins be correctly folded, modified, and delivered to specific intra- and extracellular sites. Newly synthesized membrane and secretory proteins enter a cellular sorting and distribution network during or immedialety after synthesis and are routed to specific locations inside and outside the cell. The initial compartment in this process is the endoplasmic reticulum (ER) where proteins undergo modifications such as glycosylation, disulfide bond formation, and assembly into oligomers. The modified proteins are then transported through a series of membrane-bound compartments which include the various cisternae of Golgi complex where further carbohydrate modifications occur. Transport between compartments occurs by means of vesicles that bud and fuse in a manner specific to the type of protein being transported. Once within the secretory pathway, proteins do not have to cross a membrane to reach the cell surface. Disruptions in the cellular secretory pathway have been implicated in several human diseases. In familial hypercholesterolemia the low density lipoprotein receptors remain in the ER rather than moving to the cell surface (Pathak, et al., J. Cell Biol., 106:1831-1841, 1988). Altered transport and processing of the beta-amyloid precursor protein (betaAPP) involves the putative vesicle transport protein prenesilin, and may play a role in early-onset Alzheimer disease (Levy-Lahad et al., Science, 269:973-977, 1995). Changes in the ER-derived calcium homeostasis have been associated with diseases such as cardiomyopathy, cardiac hypertrophy, myotonic dystrophy, Brody disease, Smith-McCort dysplasia and diabetes mellitus.

It is believed that the protein of the invention represents a new ER integral transmembrane protein. This protein plays probably a role in post-translational modifications of secreted and membrane proteins. Its dysregulated expression may be linked to disorders such as the above referred diseases. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID Nos: 278, 282 and 300 from positions 85 to 105, 116 to 136, 164 to 184, 187 to 207, 332 to 352, 376 to 396, 404 to 424, 465 to 485 and 499 to 519. Other preferred polypeptides of the invention are fragments of SEQ ID Nos: 278, 282 and 300 having any of the biological activity described herein.

One object of the present invention are compositions and methods of targeting heterologous polypeptides to the endoplasmic reticulum by recombinantly or chemically fusing a fragment of the proteins of the invention to an heterologous polypeptide. Preferred fragments are any fragments of the proteins of the invention, or part thereof, that may contain targeting signals for the endoplasmic reticulum such as those described in Pidoux AL, Armstrong EMBO J 1992 Apr;11(4):1583-91; Munro S, Pelham HR Cell 1987 Mar 13;48(5):899-907; Pelham HR Trends Biochem Sci 1990 Dec;15(12):483-6.

In another embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as marker proteins to selectively identify tissues, preferably salivary glands. For example, the proteins of the invention or part thereof may be used to synthesize specific antibodies using any techniques known to those skilled in the art including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry.

Moreover, antibodies to the proteins of the invention or parts thereof may be used for the detection of endoplasmic reticulum in immunochelistry for example using any techniques known to those skilled in the art including those described herein.

Protein of Seq Id No: 281 (174-10-2-F8-CS)

The protein of SEQ ID No: 281 is homologous to PET117 (SwissProt ID: Q02771). MTC is overexpressed in the brain, dystrophic muscle, fetal liver, placenta and salivary glands.

The protein of the invention, herein named MTC, presents a certain homology with the yeast PET117 protein precursor (22 % identical amino acids, 39% positive amino acids when aligned by BLASTP 2.0.9). MTC appears to be a novel member of the PET family.

Cytochrome *c* oxidase (complex IV), an enzyme complex located in the mitochondrial inner membrane, is the terminal member of the mitochondrial electron transport chain. The oxidation reaction catalyzed by cytochrome *c* oxidase is exergonic and is coupled to the translocation of protons across the membrane. This reaction provides the energy needed to drive the synthesis of ATP by the mitochondrial oxidative phosphorylation system and is essential for respiratory metabolism in aerobic eukaryotes. Cytochrome *c* oxidase is made up of as many as 13 non-identical protein subunits, of which 3 are encoded by the mitochondrial genome, and contains several prosthetic groups (including heme groups *a* and a_3).

The composition of cytochrome *c* oxidase requires that synthesis and assembly of a functional enzyme complex occur in several distinct steps, including: 1) synthesis of the protein subunits, 2) transport of the subunits from their site of synthesis to their site of function in the mitochondrial inner membrane, 3) synthesis of hemes *a* and a_3 and, 4) assembly of the subunits with each other and with the prosthetic group. A number of "accessory" genes (e.g., not encoding

protein subunits of the final assembled cytochrome c oxidase complex) are required for the production of functional cytochrome *c* oxidase (McEwen JE et al. – J Biol Chem – 1986, 261(25):11872-9). Some of these are required for the expression of mitochondrial-encoded cytochrome *c* oxidase subunits, while others are needed for the proper assembly of active cytochrome *c* oxidase.

The nuclear genes PET117 and PET 191 belong to this class of "accessory genes" required for the assembly of active mitochondrial cytochrome *c* oxidase (McEwen JE et al. – Curr Genet. – 1993, 23(1):9-14). The role of PET genes and the proteins that they encode remains obscure, although mutation experiments in *S cerevisae* have clearly shown that they are essential for the production of active cytochrome *c* oxidase (McEwen JE et al. – J Biol Chem – 1986, 261(25):11872-9) (McEwen JE et al. – Curr Genet. – 1993, 23(1):9-14).

One aspect of the subject invention provides to compositions and methods of using the nucleotide sequence of SEQ ID No: 40, or its complement, in molecular biology techniques. In one embodiment, the MTC2 sequence is encoded by clone 174-10-2-0-F8-CS. References to a polynucleotide of SEQ ID NO: 40 and polypeptide of SEQ ID NO: 281 are interchangeable with the corresponding polynucleotides of the human cDNA of clone 174-10-2-0-F8-CS and polypeptides encoded thereby. These techniques include, but are not limited to: PCR; production of recombinant MTC, or biologically active fragments thereof, generating antisense RNA and DNA, their chemical analogs and the like; hybridization probes; and chromosome gene mapping.

As is apparent to one skilled in the art, all of the non-limiting techniques listed above can be practiced with fragments of the mtc2 gene. Given the well-known nature of these techniques, the skilled artisan will be able to select an appropriate length of the mtc2 polynucleotide for use in the techniques. For recombinant expression of protein, a preferred embodiment provides the full length MTC2 gene in an expression vector.

For example, nucleotide sequence of SEQ ID No: 40 or its complement can be used to generate hybridization probes for mapping the naturally occurring genomic sequence. The sequence can be mapped to a particular chromosome or to a specific region of the chromosome using well-known techniques. These include in situ hybridization to chromosomal spreads, flow-sorted chromosomal preparations, or artificial chromosome constructions such as yeast artificial chromosomes, bacterial artificial chromosomes, bacterial P1 constructions or single chromosome cDNA libraries as reviewed in Price (Price CM – Blood Rev. – 1993, 7(2):127-34) and Trask B (Trask BJ – Trends Genet. – 1991, 7(5):149-54).

In situ hybridization of chromosomal preparations and physical mapping techniques such as linkage analysis using established chromosomal markers are invaluable in extending genetic maps that provides valuable information to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once a disease or syndrome has been crudely localized by genetic linkage to a particular genomic region, any sequences mapping to that area can represent

associated or regulatory genes for further investigation. The nucleotide sequence of the present invention can also be used to detect differences in the chromosomal location due to translocation, inversion, etc. among normal, carrier or affected individuals.

The subject invention also provides methods of using MTC polypeptides and polynucleotides encoding said polypeptides in preventing or reducing the incidence of apoptosis in cells. Dysfunctions in the mitochondrial electron transport chain result in cellular apoptosis or necrosis. In one embodiment, MTC is added to an *in vitro* culture of mammalian cells in an amount effective to reduce apoptosis. In another embodiment, cells are transfected with vectors comprising MTC polynucleotides which cause the expression of MTC peptides. MTC used in these embodiments can, optionally, contain mitochondrial targeting sequences. In another embodiment, MTC or MTC2 are encoded by clone 174-10-2-0-F8-CS.

In another embodiment, MTC polypeptides and polynucleotides encoding said polypeptides can be used in the diagnosis, treatment and/or prophylaxis of disorders associated with apoptosis or impairment of the mitochondrial respiratory electron transport chain. Polynucleotides can also be used in antisense protocols for certain disorders to impair the function of the mitochondrial electron transport chain. These disorders include, but are not limited to, immune deficiency syndromes (including AIDS); type I diabetes; pathogenic infections; cardiovascular and neurological injury; alopecia; aging; degenerative diseases such as Alzheimer's Disease, Parkinson's Disease, Huntington's disease; dystonia; Leber's hereditary optic neuropathy; schizophrenia; neonatal hepatic failure and ketoacidotic coma necrosis; and myodegenerative disorders such as "mitochondrial encephalopathy, lactic acidosis, and stroke" (MELAS), "myoclonic epilepsy ragged red fiber syndrome" (MERRF); mitochondriocytopathies, Leigh syndrome, fatal infantile cardioencephalomyopathy, ataxia; encephalopathies, aging, neurodegenerative diseases, myopathies, and cancers. As would be apparent to the routineer, these methods can be practiced with full length MTC polypeptides and polynucleotides encoding said polypeptides as well as biologically active fragments of the same which retain biological activity.

For diagnostic purposes, the expression of the protein of the invention could be investigated using any of the Northern blotting, RT-PCR or immunoblotting methods well known to those skilled in the art. For prophylaxis and/or treatment purposes SEQ ID No: 40, its complement, or fragments of either, can be used to enhance electron transport and increase energy delivery using any of the gene therapy methods known to those skilled in the art. Likewise, SEQ ID NO: MTC2, its complement, and fragments of either can be used to impair electron transport and decrease energy delivery using any of the antisense methodologies known to those skilled in the art.

Protein of SEQ ID NO:392 (145-7-2-0-G5-CS)

The protein of SEQ ID No:392, encoded by the cDNA of SEQ ID No:151, is homologous to Unc-18 proteins, also known as the STXBP or Sec-1 family. The protein of the invention is strongly expressed in the fetal kidney.

Amino acids 89 to 107 of the protein of the invention present the EMotif signature for proteins of the Sec-1 family (BlocksPlus PF00995). Furthermore, BLAST analysis (BLASTP version 2.0.9) of the amino acid sequence of the invention reveals that it is homologous to a number of proteins belonging to the Unc-18/Sec-1 family. Preferred polypeptides of the invention are those that comprise amino acids 94, 95, and/or 100, which are conserved in more than 80% of the Sec-1 family members; and/or amino acids 43, 89, and/or 97, which are conserved in more than 60% of Sec-1 family members. Other preferred polypeptides of the invention are any fragment of SEQ ID NO:392 having any of the biological activities described herein.

The normal function and organization of eukaryotic cells is dependent on transport of various vesicles that selectively shuttle membrane and cargo between distinct compartments of the secretory and endocytotic pathways. A number of key proteins involved in membrane targeting and exocytosis have been identified, and a fundamental set of interactions has been defined and placed into a model called the SNARE (Soluble N-ethylmaleimide-sensitive Attachment protein REceptor) hypothesis (Rotheman J – Nature – 1994, 372: p55-63). According to the SNARE hypothesis, vesicles dock to a target membrane through the interaction of complementary sets of vesicular (v-SNARE) and target (t-SNARE) membrane proteins. Our understanding of vesicle trafficking has, to a large extent, been facilitated by characterization of synaptic vesicles in neurons. In synaptic vesicle exocytosis, the vesicular protein synaptobrevin (also called Vesicle-Associated Membrane Protein; VAMP) is the v-SNARE, and the plasma membrane-associated protein SNAP-25 (Synaptosomal-Associated Protein of 25 kDa) and syntaxin 1 function as t-SNARE. Formation of the SNARE complex (or core complex) is followed by recruitment of the cytosolic proteins alpha, beta and gamma SNAP (Soluble N-ethylmaleimide-sensitive Attachment Protein) and NSF (N-ethylmaleimide-Sensitive Factor), which are required for membrane fusion. Proteins from two gene families have been identified as key regulators of SNARE complex assembly. These include members of the small GTP-binding family (e.g. Rabs) and the Sec-1 family. The Sec-1 gene is one of ten genes identified as essential for the final stages of protein secretion in yeast (S. cerevisae). Sec-1 homologues have been identified in the nervous system of C. elegans (Unc-18), D. melanogaster (Rop) and mammals. In mammals, the protein has been termed Mammalian homologue of the Unc-18 gene (Munc-18), rbSec1 (Rat Brain Sec1) or n-Sec1 (neural-specific Sec1).

Sec-1-related proteins are involved in the processes of vesicle targeting, docking and/or fusion. Sec-1-related proteins interact directly with the t-SNARE syntaxin, and Munc-18 has been found to interact with syntaxin isoforms 1a, 2 and 3. However, Munc-18 has not been found to be

part of the 20S SNARE/SNAP/NSF protein complex. In vitro, the binding of Munc-18 to syntaxin inhibits the interaction of syntaxin with VAMP and SNAP-25 as well as SNAP-23 (a homologue of SNAP-25) and thereby negatively regulates the formation of the synaptic SNARE fusion complex. In agreement with a negative regulatory role of Sec-1/Munc-18 proteins in neurotransmitter release are results showing that microinjections of Sec-1 into the presynaptic terminal of the giant squid synapse inhibits evoked transmitter release (Dresbach T. et al. – J Neurosci. – 1998, 18: p2923-2932). Furthermore, overexpression of Rop, Unc-18, Sec-1 and Munc-18 all result in phenotypes associated with a complete block in neurotransmitter release and/or secretion (Hosono R. et al. – J Neurochem – 1992; 58: p1517-1525; Harrison S. et al. – Neuron – 1994; 13: p555-566; Novick P. et al. – Cell – 1981; 25: p461-469; verhage M. et al. – Science – 2000; 287: p864-869). Point mutation experiments involving the Rop gene suggest that Rop is a rate-limiting regulator of exocytosis that performs both stimulatory and inhibitory functions in neurotransmission (Wu M. et al. – EMBO J – 1998; 17: p127-139). The reduction in neurotransmitter release seen after both overexpression of Munc-18 and mutations in Munc-18 homologues indicates that Sec-1 proteins not only sequester syntaxins from other proteins but also assist the syntaxins in adopting a functional conformation or facilitate interactions between syntaxins and other proteins by a chaperone-like action. The necessity of Sec1-related proteins is believed to result, in part, from their direct and high affinity interaction with members of the t-SNARE family of syntaxin proteins and from the control by this complex of a v- and t-SNARE protein interaction required for vesicle fusion.

The SNARE mechanism of exocytosis appears to be conserved both evolutionarily (most of the components have homologues in species from yeast to mammals) and functionally (each of the principal components are members of multigene families). This latter point is supported by work showing that components of this pathway are found in different cell types (neurons, neutrophiles and pancreatic beta-cells) (brumell J. et al. – J immunol – 1995; 155: p5750-5759; Zhang W et al. – J Biol Chem. – 2000 Oct 6, electronic publication).

It is believed that the protein of SEQ ID NO:392 is a member of the Unc-18/Sec-1 family, and thus plays a key role in the regulation of various processes including vesicle targeting, docking and fusion.

One embodiment of the present invention relates to the use of the protein of SEQ ID NO:392 or the cDNA of SEQ ID NO:151 or any part thereof to used to identify fetal kidney tissue and cells derived from this tissue, since the protein of the invention is strongly expressed in this tissue. In addition, the protein of the invention can be used to specifically label components of the secretory pathway within cells. Assays for the detection of cells expressing the protein of the invention, or part thereof, can be developed using techniques known to those skilled in the art. For example, the protein of the invention, or part thereof, can be used to generate antibodies or antiserum, by techniques well known to those skilled in the art. Antibodies or antiserum can also be used for quantitative analysis or detection of the protein of the invention, by methods such as

enzyme-linked immunosorbant assays (ELISA) or by any other technique known to those skilled in the art. Another possible technique involves the use of marked syntaxins, since Sec1-related proteins are known to bind to syntaxins.

In another embodiment of the present invention, the present polynuleotides and polypeptides can be used to diagnose, treat, and/or prevent any of a large number of diseases and disorders characterized by abnormal exocytosis, such as, but not limited to: allergies including hay fever, asthma, and urticaria; neurologic disorders, a number of which result from abnormal neurotransmitter secretion (for example, depression is associated with decreased serotonin secretion); autoimmune hemolytic anemia; cancers, especially hormone-dependent cancers such as those stimulated by androgens (for example, prostate cancer) or estrogens (for example, breast cancer), leukemias or lymphomas; ulcerative colitis; type 2 diabetes, which in some cases is associated with decreased insulin secretion; proliferative granulonephritis; inflammatory bowel disease; growth failure due to decreased secretion of growth hormone; multiple sclerosis; myasthenia gravis, rheumatoid and osteoarthritis; scleroderma,; Chediak-Higashi and Sjogren's syndromes; systemic lupus erythematosus; thyroiditis; toxic shock syndrome; traumatic tissue damage; viral, bacterial, fungal and protozoal infections; and other physiologic/pathologic disorders associated with induced or otherwise abnormal vesicular trafficking.

An association between the level of expression and/or activity of the present protein with the presence or absence of any condition associated with abnormal vesicular trafficking, such as any of the above-listed disorders, can readily be assessed by detecting the level of expression or activity of the protein by, e.g., Northern blot, western blot, ELISA, or any standard in vitro or in vivo assay for protein activity, and correlating the observed level or expression or activity with the presence or absence of the disorder. For those disorders found to be positively associated with the protein of the invention, a diagnostic or screening assay can be readily developed where the detection of an elevated level of protein or protein activity is indicative of the presence of the disease, or of a propensity to develop the disease. Further, any such diseases or conditions can be treated or prevented by inhibiting the expression or activity of the protein, for example by administering to a patient suffering from the disorder any inhibitor including, but not limited to, antibodies, antisense oligonucleotides, dominant negative forms of the protein, and small molecule inhibitors of protein expression or activity. Alternatively, disorders negatively associated with the protein of the invention can be diagnosed or screened for by detecting the level of the present protein or protein activity, where a decreased level of the protein or protein activity is indicative of the presence of the disease, or of a propensity to develop the disease. Such disorders negatively associated with the protein of the invention can be treated or prevented by increasing the level of the protein or protein activity, for example by administering to a patient any of a number of agents including, but not limited to, the protein itself, a polynucleotide encoding the protein, or a heterologous compound that enhances the expression or activity of the protein.

protein of SEQ ID NO:419 (internal designation 188-9-1-0-C10-CS)

The protein of SEQ ID NO:419, highly expressed in the brain and placenta, is encoded by the cDNA of SEQ ID NO:178, is localized preferentially in the endoplasmic reticulum, and is homologous to the yeast integral membrane protein SFT2p, a member of the SNARE-related family (Genbank accession number X79489). SFT2p is well conserved in C. elegans and in mice (accession numbers CAA93859 and AA790425 respectively), and plays an important role in the protein trafficking and fusion machinery of eukaryotic cells. The 159-amino-acid-long protein of the invention, which is similar in size and in membrane topology to the SFT2p protein, displays four conserved hydrophobic stretches from positions 36 to 56, 66 to 86, 98 to 118 and 122 to 142, forming a tetra-spanning membrane protein. This topology is also found in the Got1p protein, another well-conserved SNARE related protein with similar functions to those of SFT2p protein (accession number AL010285 for P. falciparum, U23521 for C. elegans) as described in Conchon et al., EMBO J., 18(14):3934-3946 (1999).

Eukaryotic proteins are synthesized within the endoplasmic reticulum (ER), are delivered from the ER to the Golgi complex for post-translational processing and sorting, and are transported from the Golgi to specific intracellular and extracellular destinations. This intracellular and extracellular movement of protein molecules is termed vesicle trafficking. Trafficking is accomplished by the packaging of protein molecules into specialized vesicles which bud from the donor organelle membrane and fuse to the target membrane (Palade, Science 189:347-358 (1975)).

Numerous proteins are necessary for the formation, targeting, and fusion of transport vesicles and for the proper sorting of proteins into these vesicles. The vesicle trafficking machinery includes coat proteins which promote the budding of vesicles from donor membranes, vesicle- and target-specific identifiers (v-SNAREs and t-SNAREs) which bind to each other and dock the vesicle to the target membrane (Nichols et al., Nature 387:199-202, 1997), and proteins which bind to SNARE complexes and initiate fusion of the vesicle to the target membrane (SNAPs).

SFT2p is a conserved yeast protein with four transmembrane domains that is resident in punctate structures corresponding to the late Golgi compartment, and which enters presumptive retrogade intra-Golgi vesicles whose fusion depends on two t-SNARE proteins Sed5p and Sft1p (Wooding and Pelham, Mol. Biol. Cell 9:2667-2680 (1998)). Its genetic interaction with Sed5p suggests that SFT2p is an additional membrane component involved in the docking or fusion process. In vivo experiments have shown that deletion of GOT1p or SFT2p alone does not affect cell growth, but repression of both of these proteins results in a significant accumulation of ER membrane, suggesting that the presence of either SFT2p or GOT1p is required for the maintenance of efficient ER-Golgi transport (Conchon et al., supra). It has also been shown that Got1p normally facilitates Sed5p-dependant fusion events, while Sft2p performs a related function in the late Golgi (Conchon et al., supra).

The etiology of numerous human diseases and disorders can be attributed to defects in the trafficking of proteins to organelles or the cell surface. For example, defects in the trafficking of membrane-bound receptors and ion channels have been implicated in cystic fibrosis (cystic fibrosis transmembrane conductance regulator; CFTR), glucose-galactose malabsorption syndrome (Na.sup.+ /glucose cotransporter), hypercholesterolemia (low-density lipoprotein (LDL) receptor), and forms of diabetes mellitus (insulin receptor). Abnormal hormonal secretion has been linked to disorders including diabetes insipidus (vasopressin), hyper- and hypoglycemia (insulin, glucagon), Grave's disease and goiter (thyroid hormone), and Cushing's and Addison's diseases (adrenocorticotropic hormone; ACTH).

Further, cancer cells secrete excessive amounts of hormones or other biologically active peptides. Disorders related to excessive secretion of biologically active peptides by tumor cells include: fasting hypoglycemia due to increased insulin secretion from insulinoma-islet cell tumors; hypertension due to increased epinephrine and norepinephrine secreted from pheochromocytomas of the adrenal medulla and sympathetic paraganglia; and carcinoid syndrome, which includes abdominal cramps, diarrhea, and valvular heart disease, caused by excessive amounts of vasoactive substances (serotonin, bradykinin, histamine, prostaglandins, and polypeptide hormones) secreted from intestinal tumors. Ectopic synthesis and secretion of biologically active peptides (peptides not expected from a tumor) includes ACTH and vasopressin in lung and pancreatic cancers; parathyroid hormone in lung and bladder cancers; calcitonin in lung and breast cancers; and thyroid-stimulating hormone in medullary thyroid carcinoma.

It is believed that the protein of SEQ ID NO:419 or part thereof is an integral membrane protein of the SNARE-related family, and more presumably is the human homologue of the yeast SFT2p protein. Thus, the protein of the invention plays a role in the secretory and endocytic pathway of eukaryotic cells through fusion and transport of vesicles from the endoplasmic reticulum to late Golgi cisternae. Preferred polypeptides of the invention are polypeptides comprising the amino acids of SEQ ID NO:419 of the four transmembrane domains from positions 36 to 56, 66 to 86, 98 to 118 and 122 to 142. Other preferred polypeptides of the invention are fragments of SEQ ID NO:419 having any of the biological activities described herein.

In one embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a new marker protein to selectively identify secretory and endocytic traffic, preferably in the endoplasmic reticulum and more preferably in the late Golgi cisternae. For example, the protein of the invention or part thereof may be detected using specific antibodies generated against the protein using any technique known to those skilled in the art. Such organelle-specific antibodies may then be used to identify cells with disrupted trafficking systems such as in differentiated tumor cells or to differentiate specific organelle types in a cell cross-section using immunochemistry. In addition, the protein of the invention can be used to specifically identify cells of the brain and/or placenta, tissues in which the protein is overexpressed.

Another embodiment of the present invention relates to methods of targeting heterologous compounds, such as polypeptides or polynucleotides, to the endoplasmic reticulum and preferentially to late Golgi vesicles by recombinantly or chemically fusing a fragment of the protein of the invention to the heterologous polypeptide or polynucleotide. Such fusion proteins may be engineered to contain a cleavage site located between a sequence encoding the protein of the invention and the heterologous protein sequence, so that the protein of the invention may be cleaved and purified away from the heterologous moiety. Preferred fragments of the protein that can be used in such applications are the four transmembrane domains or any other fragments of the protein of the invention, or part thereof, that may contain targeting signals for ER or Golgi organelles as defined in Conchon et al., supra; Wooding and Pelham, supra. Such heterologous compounds may be targeted to the secretory pathway to modulate ER-Golgi endocytic and secretory activities. In one embodiment, the protein of the invention can be used to screen peptide libraries for inhibitors of traffic activity, as detected by the accumulation of ER membranes or Golgi vesicles as described in Conchon et al., supra.

In still another embodiment, the protein of the invention is used to diagnose, prevent and/or treat any of a number of disorders in which trafficking and/or the fusion machinery is affected, including, but not limited to, endocrine, secretory, inflammatory, and gastrointestinal disorders, such as cancer, cystic fibrosis (cystic fibrosis transmembrane conductance regulator; CFTR, as well as membrane-bound receptors and ion channels associated with CFTR), glucose-galactose malabsorption syndrome (Na.sup.+ /glucose cotransporter), hypercholesterolemia (low-density lipoprotein (LDL) receptor), and forms of diabetes mellitus (insulin receptor), abnormal hormonal secretion linked to disorders including diabetes insipidus (vasopressin), hyper- and hypoglycemia (insulin, glucagon), Grave's disease and goiter (thyroid hormone), Cushing's and Addison's diseases (adrenocorticotropic hormone; ACTH), disorders related to excessive secretion of biologically active peptides by tumor cells including fasting hypoglycemia due to increased insulin secretion from insulinoma-islet cell tumors, hypertension due to increased epinephrine and norepinephrine secreted from pheochromocytomas of the adrenal medulla and sympathetic paraganglia, carcinoid syndrome, which includes abdominal cramps, diarrhea, and valvular heart disease, caused by excessive amounts of vasoactive substances (serotonin, bradykinin, histamine, prostaglandins, and polypeptide hormones) secreted from intestinal tumors. Ectopic synthesis and secretion of biologically active peptides (peptides not expected from a tumor) includes ACTH and vasopressin in lung and pancreatic cancers; parathyroid hormone in lung and bladder cancers; calcitonin in lung and breast cancers; and thyroid-stimulating hormone in medullary thyroid carcinoma.

An association between the level of expression and/or activity of the present protein with the presence or absence of any condition associated with abnormal vesicular trafficking and/or secretion, such as any of the above-listed disorders, can readily be assessed by detecting the level of expression or activity of the protein by, e.g., Northern blot, western blot, ELISA, or any standard in

vitro or in vivo assay for protein activity, and correlating the observed level or expression or activity with the presence or absence of the disorder. For those disorders found to be positively associated with the protein of the invention, a diagnostic or screening assay can be readily developed where the detection of an elevated level of protein or protein activity is indicative of the presence of the disease, or of a propensity to develop the disease. Further, any such diseases or conditions can be treated or prevented by inhibiting the expression or activity of the protein, for example by administering to a patient suffering from the disorder any inhibitor including, but not limited to, antibodies, antisense oligonucleotides, dominant negative forms of the protein, and small molecule inhibitors of protein expression or activity. Alternatively, disorders negatively associated with the protein of the invention can be diagnosed or screened for by detecting the level of the present protein or protein activity, where a decreased level of the protein or protein activity is indicative of the presence of the disease, or of a propensity to develop the disease. Such disorders that are negatively associated with the protein of the invention can be treated or prevented by increasing the level of the protein or protein activity, for example by administering to a patient any of a number of agents including, but not limited to, the protein itself, a polynucleotide encoding the protein, or a heterologous compound that enhances the expression or activity of the protein.

Cancer cells secrete excessive amounts of hormones or other biologically active peptides. Therefore, in another embodiment, antagonists or inhibitors of the protein of the invention may be administered to a subject to treat or prevent cancers by inhibiting the traffic activity in transformed cells. Any type of cancer can be treated or prevented in this way, including, but not limited to, adenocarcinoma, sarcoma, melanoma, lymphoma, and leukemia. In preferred embodiments, the cancers include cancers of glands, tissues, and organs involved in secretion or absorption, such as prostate, pancreas, lung, tongue, brain, breast, bladder, adrenal gland, thyroid, liver, uterus, ovary, kidney, testes, and organs of the gastrointestinal tract including small intestine, colon, rectum, and stomach. In a particular aspect, antibodies which are specific for the protein of the invention may be used directly as an antagonist, or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissues which express the protein of the invention. In addition, the elevated amount of the protein of the invention in tumor cells can readily be used to diagnose or screen for cancer, e.g. by measuring and comparing the level of the protein in a cell to that of a control cell using a specific antibody detected by FACS or using any other detection method known to those of skill in the art.

Protein of SEQ ID NO:297 (181-3-3-0-C9-CS)

The protein of SEQ ID No:297, encoded by the cDNA of SEQ ID NO:56, is homologous to synaptogyrin 1 (Trembl ID: Q9UGZ4). The protein of the invention is highly expressed in the brain and fetal brain, fetal liver and the testis.

The protein of SEQ ID No:297 is a splice variant of synaptogyrin 1. The splicing of the cDNA of SEQ ID NO:56 is different for exon 3: whereas exon 3 of synaptogyrin 1 is 238 base-pair long, exon 3 of SEQ ID NO:56 is 345 base-pair long. This introduces a frameshift and a stop codon. Thus, the protein of SEQ ID NO:297 is identical to synaptogyrin 1 up to and including amino acid 122, the remaining 22 amino acids are entirely different. When compared to synaptogyrin 1, the protein of the invention presents the same N-terminal domain (which is highly conserved in all synaptogyrins) and 2 of the 4 transmembrane helixes. Preferred polypeptides of the invention are those that comprise amino acids 1 to 16, which make up the N terminal cytoplasmic domain of the protein and which are highly conserved among all members of the synaptogyrin family (Kedra D et al. – Hum Genet. – 1998, 103(2):131-141). Other preferred polypeptides of the invention are those that comprise amino acids 25 to 45 and/or 68 to 88, which make up the two transmembrane alpha helixes. Thus it is believed that the protein of the invention is a member of the synaptogyrin family.

Synaptogyrins are closely related to proteins of the synaptophysin family, both of which are involved in neurotransmission and more generally in exocytosis and vesicle trafficking. Members of the synaptogyrin family include synaptogyrin 1 (with splice variants 1a, 1b and 1c), cellugyrin (synaptogyrin 2) and synaptogyrin 3. This family of proteins is also evolutionarily conserved, as homologues to human synaptogyrin 1 have been found in rats, mice, and C. elegans. Synaptogyrins and synaptophysin are among the most abundant vesicle components--together they account for more than 10% of the total vesicle membrane proteins. Although synaptogyrins do not appear to be required for exocytosis itself (apparently because synaptogyrins and synaptophysins have overlapping functions), they are essential for the normal regulation of exocytosis.

The normal function and organization of eukaryotic cells is dependent on the transport of various vesicles that selectively shuttle membrane and cargo between distinct compartments of the secretory and endocytotic pathways. A number of key proteins involved in membrane targeting and exocytosis have been identified, and a fundamental set of interactions has been defined and placed into a model called the SNARE (Soluble N-ethylmaleimide-sensitive Attachment protein REceptor) hypothesis (Rotheman J – Nature – 1994, 372: p55-63). According to the SNARE hypothesis, vesicles dock to a target membrane through the interaction of complementary sets of vesicular (v-SNARE) and target (t-SNARE) membrane proteins. Our understanding of vesicle trafficking has, to a large extent, been facilitated by characterization of synaptic vesicles in neurons. In synaptic vesicle exocytosis, the vesicular protein synaptobrevin and synaptogyrin (also called Vesicle-Associated Membrane Protein; VAMP) are the v-SNARE, and the plasma membrane-associated protein SNAP-25 (Synaptosomal-Associated Protein of 25 kDa) and syntaxin 1 function as t-SNARE. Formation of the SNARE complex (or core complex) is followed by recruitment of the cytosolic proteins alpha, beta and gamma SNAP (Soluble N-ethylmaleimide-sensitive Attachment Protein) and NSF (N-ethylmaleimide-Sensitive Factor), which are required for membrane fusion.

In transfected PC12 cells, synaptogyrin 1 and synaptophysin 1 are as effective as tetanus toxin light chain in inhibiting exocytosis (Sugita S. et al. - J Biol Chem. - 1999, 274(27):18893-901), suggesting that these proteins are strong regulators of exocytosis. More recently, synaptogyrins have been found to have an essential function in synaptic plasticity (Janz R. et al - Neuron. - 1999, 24(3):687-700).

The etiology of numerous human diseases and disorders can be attributed to defects in the trafficking of proteins to organelles or the cell surface. For example, defects in the trafficking of membrane-bound receptors and ion channels have been implicated in cystic fibrosis (cystic fibrosis transmembrane conductance regulator; CFTR), glucose-galactose malabsorption syndrome (Na.sup.+ /glucose cotransporter), hypercholesterolemia (low-density lipoprotein (LDL) receptor), and forms of diabetes mellitus (insulin receptor). Abnormal hormonal secretion has been linked to disorders including diabetes insipidus (vasopressin), hyper- and hypoglycemia (insulin, glucagon), Grave's disease and goiter (thyroid hormone), and Cushing's and Addison's diseases (adrenocorticotropic hormone; ACTH).

Further, cancer cells secrete excessive amounts of hormones or other biologically active peptides. Disorders related to excessive secretion of biologically active peptides by tumor cells include: fasting hypoglycemia due to increased insulin secretion from insulinoma-islet cell tumors; hypertension due to increased epinephrine and norepinephrine secreted from pheochromocytomas of the adrenal medulla and sympathetic paraganglia; and carcinoid syndrome, which includes abdominal cramps, diarrhea, and valvular heart disease, caused by excessive amounts of vasoactive substances (serotonin, bradykinin, histamine, prostaglandins, and polypeptide hormones) secreted from intestinal tumors. Ectopic synthesis and secretion of biologically active peptides (peptides not expected from a tumor) includes ACTH and vasopressin in lung and pancreatic cancers; parathyroid hormone in lung and bladder cancers; calcitonin in lung and breast cancers; and thyroid-stimulating hormone in medullary thyroid carcinoma.

In one embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a new marker protein to selectively identify secretory and endocytic traffic, preferably in the endoplasmic reticulum and more preferably in the late Golgi cisternae. For example, the protein of the invention or part thereof may be detected using specific antibodies generated against the protein using any technique known to those skilled in the art. Such organelle-specific antibodies may then be used to identify cells with disrupted trafficking systems such as in differentiated tumor cells or to differentiate specific organelle types in a cell cross-section using immunochemistry. In addition, the protein of the invention can be used to specifically identify cells of the brain, fetal brain, fetal liver and the testis, tissues in which the protein is overexpressed.

Another embodiment of the present invention relates to methods of targeting heterologous compounds, such as polypeptides or polynucleotides, to the components of the secretory machinery by recombinantly or chemically fusing a fragment of the protein of the invention to the heterologous

polypeptide or polynucleotide. Such fusion proteins may be engineered to contain a cleavage site located between a sequence encoding the protein of the invention and the heterologous protein sequence, so that the protein of the invention may be cleaved and purified away from the heterologous moiety. Such heterologous compounds may be targeted to the secretory pathway to modulate ER-Golgi endocytic and secretory activities. In one embodiment, the protein of the invention can be used to screen peptide libraries for inhibitors of traffic activity, as detected by the accumulation of ER membranes or Golgi vesicles as described in Conchon et al., supra.

In still another embodiment, the protein of the invention is used to diagnose, prevent and/or treat any of a number of disorders in which trafficking and/or the fusion machinery is affected, including, but not limited to, endocrine, secretory, inflammatory, and gastrointestinal disorders, such as cancer, cystic fibrosis (cystic fibrosis transmembrane conductance regulator; CFTR, as well as membrane-bound receptors and ion channels associated with CFTR), glucose-galactose malabsorption syndrome (Na.sup.+ /glucose cotransporter), hypercholesterolemia (low-density lipoprotein (LDL) receptor), and forms of diabetes mellitus (insulin receptor), abnormal hormonal secretion linked to disorders including diabetes insipidus (vasopressin), hyper- and hypoglycemia (insulin, glucagon), Grave's disease and goiter (thyroid hormone), Cushing's and Addison's diseases (adrenocorticotropic hormone; ACTH), disorders related to excessive secretion of biologically active peptides by tumor cells including fasting hypoglycemia due to increased insulin secretion from insulinoma-islet cell tumors, hypertension due to increased epinephrine and norepinephrine secreted from pheochromocytomas of the adrenal medulla and sympathetic paraganglia, carcinoid syndrome, which includes abdominal cramps, diarrhea, and valvular heart disease, caused by excessive amounts of vasoactive substances (serotonin, bradykinin, histamine, prostaglandins, and polypeptide hormones) secreted from intestinal tumors. Ectopic synthesis and secretion of biologically active peptides (peptides not expected from a tumor) includes ACTH and vasopressin in lung and pancreatic cancers; parathyroid hormone in lung and bladder cancers; calcitonin in lung and breast cancers; and thyroid-stimulating hormone in medullary thyroid carcinoma.

An association between the level of expression and/or activity of the present protein with the presence or absence of any condition associated with abnormal vesicular trafficking and/or secretion, such as any of the above-listed disorders, can readily be assessed by detecting the level of expression or activity of the protein by, e.g., Northern blot, western blot, ELISA, or any standard in vitro or in vivo assay for protein activity, and correlating the observed level or expression or activity with the presence or absence of the disorder. For those disorders found to be positively associated with the protein of the invention, a diagnostic or screening assay can be readily developed where the detection of an elevated level of protein or protein activity is indicative of the presence of the disease, or of a propensity to develop the disease. Further, any such diseases or conditions can be treated or prevented by inhibiting the expression or activity of the protein, for example by administering to a patient suffering from the disorder any inhibitor including, but not limited to,

antibodies, antisense oligonucleotides, dominant negative forms of the protein, and small molecule inhibitors of protein expression or activity. Alternatively, disorders negatively associated with the protein of the invention can be diagnosed or screened for by detecting the level of the present protein or protein activity, where a decreased level of the protein or protein activity is indicative of the presence of the disease, or of a propensity to develop the disease. Such disorders that are negatively associated with the protein of the invention can be treated or prevented by increasing the level of the protein or protein activity, for example by administering to a patient any of a number of agents including, but not limited to, the protein itself, a polynucleotide encoding the protein, or a heterologous compound that enhances the expression or activity of the protein.

Cancer cells secrete excessive amounts of hormones or other biologically active peptides. Therefore, in another embodiment, antagonists, inhibitors, or other modulators of the protein of the invention may be administered to a subject to treat or prevent cancers by inhibiting the traffic activity in transformed cells. Any type of cancer can be treated or prevented in this way, including, but not limited to, adenocarcinoma, sarcoma, melanoma, lymphoma, and leukemia. In preferred embodiments, the cancers include cancers of glands, tissues, and organs involved in secretion or absorption, such as prostate, pancreas, lung, tongue, brain, breast, bladder, adrenal gland, thyroid, liver, uterus, ovary, kidney, testes, and organs of the gastrointestinal tract including small intestine, colon, rectum, and stomach. In a particular aspect, antibodies which are specific for the protein of the invention may be used directly as an antagonist, or indirectly as a targeting or delivery mechanism for bringing a pharmaceutical agent to cells or tissues which express the protein of the invention.

In addition, the present protein can be used to diagnose, treat, and prevent any neurological or psychiatric disorder or condition associated with abnormal neurotransmitter release, such as depression, which is associated with decreased serotonin secretion, or any neurological function, e.g. memory, which could be enhanced or otherwise modulated by altering the quantity, frequency, or any other property of neurotransmitter release in one or more cell types in the nervous system.

Proteins of SEQ ID NOs:247 and 246 (internal designations 105-031-2-0-D3-CS and 105-031-1-0-A2-CS)

The protein of SEQ ID NOs:247 and 246, encoded by the cDNAs of SEQ ID NOs:6 and 5, respectively, are overexpressed in liver, pancreas, and prostate. The proteins of the invention are strongly homologous to the human membrane-bound protein PRO836 (GENSEQP accession number: W63687), and to the human secreted protein 7 (GENSEQP accession number: Y57941). The proteins of the invention also share homology with the chaperone-associated protein, SLS1p, found in yeast Yarrowia lipolytica (GENPEPT accession number Z50154), having 27% identity from amino-acids 68 to 340 of protein of SEQ ID No:247. In addition, the proteins of SEQ ID NOs:247 and 246 share homology with two Hsp70 family proteins, Hsp-binding protein 1 found in

mice (GENSEQP accession number: Z50154), and human species (GENPEPT accession number: AF093420), and Hsp-binding protein 2 found in human species (GENPEPT accession number: AF187859).

The proteins of the invention are related to a yeast lumen protein of the endoplasmic reticulum, SLS1p. This protein acts in the preprotein translocation process, interacting directly with translocating polypeptides to facilitate their transfer and/or help their folding in the endoplasmic reticulum (Boisrame et al. J Biol Chem 1996; 271:11668-75). In addition, Sls1p is believed to act as a cofactor of the chaperon protein Kar2 (Boisrame et al. J Biol Chem 1998; 273:30903-8 ; Kabani et al. Gene 2000; 241:309-15). Thus, the proteins of the invention are presumed to have similar cellular functions as those of chaperones. Such functions include a number of cellular processes, such as protein folding, disassembly of oligomeric protein structures, regulation of apoptosis, protein degradation, protein translocation in the endoplasmic reticulum, and antigen-presentation (Bukau et al. Cell 1998; 92:351-66). Chaperones are also involved in a number of disorders, especially autoimmune diseases such as type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus, Sjogren syndrome, and mixed connective tissue disease (Feige et al. EXS 1996; 77:359-73; Feili-Hariri et al. J Autoimmun 2000; 14:133-42). Chaperones are also involved in various disorders including tuberculosis and leprosy (Zugel et al. Clin Microbiol Rev 1999; 12:19-39), neurogenerative disorders such as Alzheimer and Parkinson diseases (Yoo et al. J Neural Transm Suppl 1999; 57:315-22), and malignant disorders (Csermely et al. Pharmacol Ther 1998; 79:129-68). In addition, a growing body of evidence suggests the involvement of the Hsp60 chaperone in the development of atherosclerosis (Xu et al. Circulation 2000; 102:14-20). Thus, the present proteins, which are presumed to be co-factors of a chaperon as summarized above, are believed to have analogous cellular functions and to be involved in similar pathological processes.

In one embodiment, the present invention provide methods of using the present proteins to identify specific cell types in vitro and in vivo. For example, as chaperone proteins are often upregulated in response to cellular stress, the detection of cells expressing elevated levels of the proteins provides a tool for detecting cells under stress. As cellular stress has been implicated in a number of disorders, such as cardiovascular disorders, neurodegenerative disorders, and cancer, the ability to detect such stress thus provides a diagnostic or screening tool for such conditions. In addition, the present polypeptides and polynucleotides can be used to identify liver, pancreas, and prostate tissues, and cells derived from these tissues. The ability to specifically visualize such tissues and cells is useful for a number of applications, including to determine the origin or identity of, *e.g.* cancerous cells, as well as to facilitate the identification of particular cells and tissues for, *e.g.* the evaluation of histological slides.

In addition, the present polypeptides and polynucleotides can be used to develop diagnostic and screening assays for diseases characterized by an abnormal level or activity of the protein of SEQ ID NOs:247 and 246. Such disorders include, but are not limited to, infectious diseases,

neurogenerative disorders such as Alzheimer's and Parkinson's diseases, schizophrenia, alopecia, aging, atherosclerosis, malignant disorders of various types, and autoimmune diseases including type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus, Sjogren syndrome, and mixed connective tissue disease. Such assays can be performed using any biological sample, such as serum or plasma.

In still another embodiment, the proteins of the invention or part thereof can be used to prevent cells from undergoing apoptosis. Specifically, as chaperone proteins have been shown to protect cells from apoptosis, any method of increasing the level or activity of the present protein can be used to prevent cells from undergoing apoptosis, in vitro or in vivo. For example, a polynucleotide encoding a protein of SEQ ID NO:247 or 246, or any fragment or derivative thereof, can be introduced into cells, e.g. in a vector, wherein the protein is expressed in the cells. Alternatively, a protein of SEQ ID NO:247 or 246 itself can be administered to cells, preferably in a formulation that leads to the internalization of the protein by the cells. Also, any compound that increases the expression or activation of the proteins within the cells can be administered. Preventing cells from undergoing apoptosis can be used for any of a large number of purposes, including, but not limited to, to prevent the death of cells being grown in culture, to prevent in a patient the apoptosis associated with any of a number of disorders, or to prevent apoptosis in cells of a patient undergoing a treatment that increases the level of cellular stress, such as chemotherapy.

In another embodiment, inhibiting the proteins of the invention can be used to induce apoptosis in undesired cells. Such inhibition can be accomplished in any of a number of ways, including, but not limited to, using antibodies, antisense sequences, dominant negative forms of the protein, or small molecule inhibitors of the expression or activity of the proteins. Such induction of apoptosis can be used to eliminate any undesired cells, for example cancer cells, in a patient. Preferably, such inhibitors are targeted specifically to the undesired cells in the patient.

In another embodiment, various disorders can be treated, attenuated and/or prevented by a protein of SEQ ID NOs:247 or 246, or part thereof, or any other compound that can affect the level or activity of the proteins such as nucleic acids, antibodies, or chemical substances. In a preferred embodiment, proteins or other compounds directed to the proteins of the invention can be used to treat or prevent disorders in which the activity or level of the proteins of SEQ ID NO:247 or 246 is unbalanced. Such diseases include, but are not limited to, infectious diseases, neurogenerative disorders as Alzheimer and Parkinson diseases, schizophrenia, alopecia, aging, atherosclerosis, malignant disorders of various types, and autoimmune diseases including type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus, Sjogren syndrome, mixed connective tissue disease, malignant disorders, autoimmune and any other neurodegenerative disorder. In another embodiment, the proteins of SEQ ID NO:247 or 246 or part thereof can be used as vaccines for various disorders including, but not limited, to cancer (Wang et al. Immunol Invest 2000;29:131-7),

tuberculosis (Silva et al. Microbes Infect 1999;1:429-35), diabetes (Int Immunol 1999;11:957-66), and atherosclerosis (Xu et al. Arterioscler Thromb 1992;12:789-99).

Protein of SEQ ID NO:389 (internal designation 109-003-1-0-G4-CS)

The protein of SEQ ID NO:389 is encoded by the cDNA of SEQ ID NO:148. Accordingly, it will be appreciated that all characteristics and uses of the polypeptide of SEQ ID NO:389 described throughout the present application also pertain to the polypeptide encoded by the human cDNA of clone 109-003-1-0-G4-CS. In addition, it will be appreciated that all characteristics and uses of the nucleic acid of SEQ ID NO:148 described throughout the present application also pertain to the human cDNA of clone 109-003-1-0-G4-CS. The protein of SEQ ID NO:389 is highly homologous to two human proteins encoded by genes listed in Genbank under accession numbers AF143723 and AF112210, the disclosures of which are incorporated herein by reference in their entireties.

The polypeptide encoded by Genbank accession numbers AF143723 and AF112210 belong to the Hsp70 protein family (even though one of them has erroneously been attributed to the related Hsp60 family). Many genes encoding "Hsps" (heat shock proteins) have been cloned and sequenced, including, for example, human hsp70 (GenBank Accession Nos. M11717 and M15432; see also Hunt and Morimoto, 1985, Proc. Natl. Acad. Sci. USA 82: 6455-6459, the disclosures of which are incorporated herein by reference in their entireties), human hsp90 (GenBank Accession No. X15183; see also Yamazaki et al., 1989, Nucleic Acids Res. 17: 7108, the disclosures of which are incorporated herein by reference in their entireties), and human gp96 (GenBank Accession No. M33716; see also Maki et al., 1990, Proc. Natl. Acad. Sci. USA 87: 5658-5662, the disclosures of which are incorporated herein by reference in their entireties).

The protein of SEQ ID NO:389 and the two homologs mentioned above are actually closer to yeast members of the family than to the human Hsp70, which makes the corresponding genes previously unidentified human members of the family. Both the Pfam and Prosite Hsp70 signatures (respectively the "HSP70" Pfam model from amino acid position 3 to 509 and the PS01036 Prosite motif from position 332 to 346) are recognized within the protein of SEQ ID NO: 389. The protein of SEQ ID NO:389 differs from the protein encoded by AF112210 at amino-acid positions 282, 312 and 326, and from the protein encoded by AF143723 at amino acid position 15 and 326.

Heat shock proteins are a family of molecular chaperone proteins which have long been known to play essential roles in a multitude of intra-and intercellular processes, including protein synthesis and folding, vesicular trafficking, and antigen processing and presentation. Hsps are among the most highly conserved proteins known, and carry out many of their regulatory activities via protein-protein interactions. Historically they were identified by induction under conditions of stress, during which they are now known to provide an essential action of preventing aggregation

and assisting refolding of misfolded proteins. The major stress proteins accumulate to very high levels in stressed cells but occur at low to moderate levels in cells that have not been stressed.

Hsp70 is one member of the heat shock protein family. (Milner, C. M. and Campbell, R. D. Immunogenetics 32: 242-251 (1990); Genbank Accession No. M59828, the disclosures of which are incorporated herein by reference in their entireties). The 70kD heat shock protein is a highly conserved, ubiquitous protein involved in chaperoning proteins to various cellular organelles. Contrary to other members of the Hsp family, it is highly inducible in mammals. Although Hsp70 is barely detectable at normal temperatures, it becomes one of the most actively synthesized proteins in the cell upon heat shock (Welch et al., 1985, J. Cell. Biol. 101:1198-1211, the disclosure of which is incorporated herein by reference in its entirety). In contrast, the Hsp90 and Hsp60 proteins are abundant at normal temperatures in most, but not all, mammalian cells and are further induced by heat (Lai et al., 1984, Mol. Cell. Biol. 4:2802-10; van Bergen en Henegouwen et al., 1987, Genes Dev., 1:525-31, the disclosures of which are incorporated herein by reference in their entireties). Furthermore the Hsp70 proteins act as monomers whereas the functionally related Hsp60 proteins are associated in vivo within large double ring assemblies of nearly a million daltons. The various actions of the Hsps all rely basically on their ability to complex polypeptide segments, preferrably hydrophobic, and to stabilize them in an extended conformation in an ATP-dependent manner. The complexed polypeptides can be antigenic peptides (in which case the Hsps help directing them to the major histocompatibility complexes for presentation) or misfolded proteins which are facilitated to adopt the proper conformation by repeated cycles of binding to Hsps followed by release/refolding (see Bukau, B. and Horwich L., 1998, Cell 92: 351-366, the disclosure of which is incorporated herein by reference).

On the basis of the above information, it is believed that the protein of SEQ ID NO:389 is a member of the human Hsp70 family. Accordingly, the protein of SEQ ID NO:389 may play a role in protein synthesis/folding, cellular trafficking , antigen processing, the cellular stress response and the immune response in immuno-competent cell types. Additional information regarding the protein of SEQ ID NO:389 may be obtained by performing a binding assay with a consensus Hsp70 substrate using the methods described in Rüdiger et al., 1997, EMBO J. 16, 1501-1507, the disclosure of which is incorporated herein by reference in its entirety.

One embodiment of the present invention relates to methods of using the protein of SEQ ID NO:389 or fragments comprising at least 5, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, or 200 consecutive amino acids thereof, or fragments having a desired biological activity as a stabilizing adjuvant to slow down protein degradation, boost the yields of recombinant proteins or regenerate denatured proteins. In such an embodiment, the protein of SEQ ID NO:389 of fragment thereof is mixed with a composition comprising the protein for which it is desired to slow down degradation, boost yield, or regenerate denatured proteins under conditions which facilitate the desired result.

For example, numerous commercial assay kits commonly used by those skilled int the arts of molecular biology and biochemistry depend on the biological properties of proteins (mostly enzymes) which can be very short-lived in vitro due to the low stability of those proteins. An example is described in Eur. Patent DE4124286, the disclosure of which is incorporated herein by reference in its entirety, wherein the low intrinsic stability of test solutions used in optical tests is increased by addition of chaperone proteins, thus making the test more sensitive.

The protein of SEQ ID NO:389 may also be used to increase the yield or activity of recombinant proteins. In recombinant DNA technology, a major unsolved problem is the solubility and biological activity of the recombinantly overexpressed protein in a host, especially a bacterial or yeast host. Many eukaryotic proteins, especially the secreted ones, require for correct folding a specific cellular machinery which is lacking in bacterial hosts such as *E. coli* or becomes insufficient in mammalian/yeast cells due to high expression of the protein. The ability of the protein of SEQ ID NO:389 or fragments thereof to ensure proper folding of recombinant proteins may be utilized as follows. The protein of SEQ ID NO:389, may be coexpressed with the recombinant protein in bacterial or eukaryotic hosts to cause the hosts to express the heterologous proteins or polypeptides in a form having increased solubility and/or biological activity. For example, the protein of SEQ ID NO:389 or fragments thereof may be used in the methods described in PCT application WO 93/25681, the disclosure of which is incorporated herein by reference in its entirety. Alternatively the protein of SEQ ID NO:389 or fragments thereof may be exogeneously added to the cell cultures as described in PCT application WO 00/08135, the disclosure of which is incorporated herein by reference in its entirety. Indeed PCT application WO 00/31113, the disclosure of which is incorporated herein by reference in its entirety, shows that when added exogenously to cells, Hsp70 is readily imported into both cytoplasmic and nuclear compartments. Preparation and purification of the protein of SEQ ID NO:389 or fragments thereof may be carried out as described in Patent US-6,007,821, the disclosure of which is incorporated herein by reference in its entirety.

The protein of SEQ ID NO:389 or fragments thereof. may further be used to regenerate denatured proteins. Recombinantly expressed proteins with poor biologival activity are routinely denatured with a potent denaturing agent, such as guanidine hydrochloride, followed by refolding by dilution with a large amount of a diluent to reduce the concentration of the denaturing agent. However, this method often results in a poor refolding rate which may be significantly increased by addition of a cocktail of chaperone proteins in a fashion similar to that described for Hsp60in Eur. Patent EP0650975, the disclosure of which is incorporated herein by reference in its entirety. The advantage of using a cocktail of chaperone proteins is to accommodate differences in binding specificity of the Hsp different families and the different members within each family. For instance, vertebrate actin is efficiently folded by the chaperonine of the eukaryotic cytosol (Gao et al., 1992, Cell 69:1043-1050, the disclosure of which is incorporated herein by reference in its entirety) but

not at all by Hsp60 (Tian et al., 1995, Nature 375:250-253, the disclosure of which is incorporated herein by reference in its entirety).

Another embodiment of the present invention relates to the use of the protein of SEQ ID NO:389 or fragments thereof to deliver heterologous compounds (proteins, peptides, or DNA) to specific cellular compartments, preferably the cytoplasm and the nucleus. If desired, the protein of SEQ ID NO:389 or a fragment thereof may be fused to the heterologous compound. For example, the protein of SEQ ID NO:389 or fragments thereof may be used to chaperone compounds into cells using the methods described in PCT application WO 00/31113, the disclosure of which is incorporated herein by reference in its entirety. In the methods described in WO 00/3113, Hsp70 was used to deliver NF-KB, a key transcriptional regulator of inflammatory responses, into the nuclear compartment. It was shown that a fusion protein composed of a Cterminal Hsp70 peptide and amino acids 37-409 of the p50 subunit of NF-KB was directed into the nucleus of cells, could bind DNA specifically, and activated kappa Ig expression and TNFa production.

In one embodiment of the present invention, the protein of SEQ ID NO:389 or a fragment thereof may be used in human therapy as a modulator of immune response. Disease states which may be treated by Hsp70, fragments thereof, and/or Hsp70 complexes of the present invention include transplant rejection (see US5,891,653, the disclosure of which is incorporated herein by reference in its entirety) and autoimmune diseases, such as insulin dependent diabetes mellitus, rheumatoid arthritis, multiple sclerosis, juvenile diabetes, asthma, and inflammatory bowel disease, as well as inflammatory diseases, cancer, viral replication diseases and vascular diseases as described in the following patents, each of which is incorporated herein by reference in its entirety: US6,007,821; WO 00/31113; WO 99/18801 (treatment of auto-immune diseases), US6,017,540; US6,017,544; AU3425899; WO 99/54464; US5,837,251; US5,830,464; WO 98/34642; WO 98/34641; US5,750,119; WO 97/10001; WO 96/10411(cancer treatment); DE19813760, DE19813759 (both autoimmune disease and cancer).

The protein of SEQ ID NO:389 or fragments thereof may also be used to treat or ameliorate autoimmune disease. In this embodiment, compositions of complexes of heat shock/stress proteins (including, but not limited to the protein of SEQ ID NO:389) are administered to an individual suffering from an autoimmune disease. The complexes may be comprised of the protein of SEQ ID NO:389 or fragments thereof alone or may include other heat shock/stress proteins. In one embodiment, the protein of SEQ ID NO:389 or a fragment thereof is bound noncovalently to antigenic molecules and administered to individuals suffering from autoimmune disease to suppress the autoimmune response. Alternatively, compositions comprising the protein of SEQ ID NO:389 or fragments thereof in an un-complexed form (i.e., free of antigenic molecules) may also be administered to an individual suffering from autoimmune disease to suppress the immune response (see Patent US6,007,821, the disclosure of which is incorporated herein by reference in its entirety).

The ability of stress proteins to chaperone the antigenic peptides of the cells from which they are derived allows them to be used to isolate the antigenic peptides expressed in a tumor. In this embodiment of the present invention, complexes comprising the protein of SEQ ID NO:389 or fragments thereof and an antigenic peptide expressed by the tumor are isolated. The isolated complexes are administered back to the individual from which they were obtained in order to elicit an immune response against the tumor. Accordingly, this approach circumvents the necessity of isolating and characterizing specific tumor antigens and enables the skilled artisan to readily prepare immunogenic compositions effective against a tumor in an individual (see Patent US6,017,544, the disclosure of which is incorporated herein by reference in its entirety).

The protein of SEQ ID NO:389 may also be used to diagnose bladder cancer. The segment of the protein of SEQ ID NO:389 extending between amino acid positions 1 through 187 is more than 99% identical to a polypeptide which is linked to bladder cancer. (See Eur. Patent DE19818620, the disclosure of which is incorporated herein by reference in its entirety). The 187 amino-acid long polypeptide described in DE19818620 was identified as the partial product of the only gene for which expression was significantly altered in a bladder tumour compared to a healthy bladder. In another embodiment of the present invention, the protein of SEQ ID NO:389 or a fragment thereof thereof may be used to diagnose disorders associated with altered intercellular communication or secretion. In such techniques, the level of the protein of SEQ ID NO:389 in an individual is measured using techniques such as those described herein. The level of the protein of SEQ ID NO:389 in the individual is compared to the level in normal individuals. An altered level of the protein of SEQ ID NO:389 relative to normal individuals suggests that the individual is suffering from bladder cancer. The level of the protein of SEQ ID NO:389 present in the individual may determined by contacting a sample from the individual with an antibody directed against the polypeptide of SEQ ID NO:389 . Alternatively, the level of the protein of SEQ ID NO:389 in the individual may be measured by determining the level of RNA encoding the protein of SEQ ID NO:389 in the sample. RNA levels may be measured using nucleic acid arrays or using techniques such as in situ hybridization, Northern blots, dot blots or other technques familiar to those skilled in the art. If desired, an amplification reaction, such as a PCR reaction, may be performed on the nucleic acid sample prior to analysis. The level of RNA in the sample is compared to RNA levels in normal individuals to determine whether the individual is suffering from bladder cancer.

Antibodies against the protein of the protein of SEQ ID NO:389 or nucleic acid probes complementary to the sequence encoding the protein of SEQ ID NO:389 may also be used as a prognosis of tumor recurrence in breast as described in Patent US Patent No.: 5,188,964, the disclosure of which is incorporated herein by reference in its entirety. As described in U.S. Patent No. 5,188,964, specific levels of the stress response proteins (including Hsp70) were identified, above which the probability of tumor recurrence is highly signficant. Accordingly, the levels of the protein of SEQ ID NO:389 or RNA encoding the protein of SEQ ID NO:389 may be determined

from in a sample from an individual who has experienced a breast tumor in the past. Protein or RNA levels may be measured as described herein. If the protein or RNA levels exceed the levels above which tumor occurrence is likely, an appropriate course of treatment may be initiated.

In another embodiment of the present invention, the protein of SEQ ID NO:389 may be used to promote tissue repair and/or increase cell survival in stress conditions such as hypoxy, oxidative stress, genotoxic agents and more generally harmful conditions leading to programmed cell death. The beneficial effect is produced either by protecting the cell proteins from premature denaturation/degradation or by directly inhibiting a signal transduction pathway leading to programmed cell death (Gabai VL. et al., 1998, FEBS Lett. 438:1-4, the disclosure of which is incorporated herein by reference in its entirety). Those conditions include but are not limited to infarction, heart surgery, stroke, neurodegenerative diseases, epilepsy, trauma, atherosclerosis, restenosis after angioplasty, and nerve damage. For example, it is known that hypoxic stress is a signal that increases the amount of Hsp70 in cardiac tissue, whereupon Hsp70 helps cells survive by binding to partially denatured proteins and assisting in the refolding of these proteins into more stable native structures. Such assistance would be extremely important in providing protection to the heart during periods of hypoxia such as during an infarct or during surgery when blood flow to the heart may be temporarily halted. Several groups have also shown that overproduction of Hsp70 leads to protection in several different models of nervous system injury (reviewed in Midori AY et al., 1999, Mol. Med. Today, 5:525-31, the disclosure of which is incorporated herein by reference in its entirety). Therapeutic methods for administering the protein of SEQ ID NO:389 or a fragment thereof include but are not limited to those disclosed in Patent WO 00/23093, the disclosure of which is incorporated herein by reference in its entirety.

Accordingly, it may be desirable to increase or decrease the level of the protein of SEQ ID NO:389 in an individual having a condition resulting from an increased or decreased level of the protein. In such embodiments, the protein of SEQ ID NO:389 , or a fragment thereof, is administered to an individual in whom it is desired to increase or decrease any of the foregoing activities. The protein of SEQ ID NO:389 or fragment thereof may be administered directly to the individual or, alternatively, a nucleic acid encoding the protein of SEQ ID NO:389 or a fragment thereof may be administered to the individual. Alternatively, an agent which increases the activity of the protein of SEQ ID NO:389 may be administered to the individual. Such agents may be identified by contacting the protein of SEQ ID NO:389 or a cell or preparation containing the protein of SEQ ID NO:389 with a test agent and assaying whether the test agent increases the activity of the protein. For example, the test agent may be a chemical compound or a polypeptide or peptide.

Alternatively, the activity of the protein of SEQ ID NO:389 may be decreased by administering an agent which interferes with such activity to an individual. Agents which interfere with the activity of the protein of SEQ ID NO:389 may be identified by contacting the protein of

SEQ ID NO:389 or a cell or preparation containing the protein of SEQ ID NO:389 with a test agent and assaying whether the test agent decreases the activity of the protein. For example, the agent may be a chemical compound, a polypeptide or peptide, an antibody, or a nucleic acid such as an antisense nucleic acid or a triple helix-forming nucleic acid.

Protein of SEQ ID NO:250 (internal designation 105-053-4-0-E8-CS)

The protein of SEQ ID NO:250 is encoded by the cDNA of SEQ ID NO:9. It will be appreciated that all characteristics and uses of the polypeptide of SEQ ID NO:250 described throughout the present application also pertain to the polypeptide encoded by the human cDNA of clone 105-053-4-0-E8-CS. In addition, it will be appreciated that all characteristics and uses of the nucleic acid of SEQ ID NO:9 described throughout the present application also pertain to the human cDNA of clone 105-053-4-0-E8-CS. The protein of SEQ ID NO:250 is found in prostate and exhibits extensive homologies to stretches of pancreatic zymogen granule membrane protein GP2 (Glycoprotein-2). In particular, the protein of SEQ ID NO:250 exhibits homologies to the GP2 proteins of human (SWISS-PROT accession number P55259, the disclosure of which is incorporated herein by reference in its entirety), rat (SWISS-PROT accession number P19218, the disclosure of which is incorporated herein by reference in its entirety) and dogs (SWISS-PROT accession number P25291, the disclosure of which is incorporated herein by reference in its entirety). In fact, the amino acid sequence of SEQ ID NO:250 is completely identical to those of human GP2 sequences except that the protein of SEQ ID NO:250 is missing amino acids 62 to 484 from the human GP2 sequence. The protein of SEQ ID NO:250 contains two hydrophobic regions, namely the N-terminal signal peptide (amino acid residues 8-28) and the C-terminal transmembrane domain (amino acid residues 91-111).

GP2 (Glycoprotein-2) is the major membrane glycoprotein of secretory zymogen granule (ZG) membranes within pancreatic acinar cells (Fukuoka et al. 1990 Nuc. Acids Res., 18:5900; Fukuoka et al. 1991 Proc. Natl. Acad. Sci., USA, 88:2898-2902; Fukuoka et al. 1992 Proc. Natl. Acad. Sci. USA, 89:1189-1193; Freedman, et al. 1993 Eur. J. Cell Biol. 61:229-238; Scheele et al. 1993 Pancreas :139-149; Freedman et al. 1994 Annals N.Y. Acad. Sci. 713:199-206, the disclosures of which are incorporated herein by reference in their entireties). GP2 homologues are also widely distributed among diverse epithelial tissues known to possess regulated secretory processes, including parotid, submandibular gland, stomach, liver and lung (Fukuoka et al. 1992 Proc. Natl. Acad. Sci. USA, 89:1189-1193).

In addition to ZG membranes, GP2 is also located in pancreatic acinar cells in rough endoplasmic reticulum, Golgi, trans-Golgi components, condensing vacuoles, apical plasma membranes (APM), basolateral plasma membranes (BPM), and within ZGs and acinar lumina (Scheele et al., 1994 Pancreas 9:139-149). GP2 is linked to the membrane of the ZG via a glycosylphosphatidyl inositol-anchor (GPI-anchor) (Fukuoka et al. 1991 Proc. Natl. Acad. Sci.

USA, 88:2898-2902; Lebel and Beattie 1988 Biochem. Biophys. Res. Comm. 254:1189-93, the disclosures of which are incorporated herein by reference in their entireties) and forms complexes, usually tetrameric complexes, below a pH of about 6.5.

During assembly of secretory granules within the trans-Golgi network (TGN), the low pH of the TGN causes formation of GP2 complexes. These complexes bind to proteoglycans (PG), forming a fibrillar GP2/PG meshwork on the lumenal surface of the ZG. The GP2/PG matrix may function in membrane sorting within the TGN, assembly of ZG membranes, inactivation of ZG membranes during granule storage, and regulation of ZG membrane trafficking at the apical plasma membrane. The GP2/PG matrix may also protect the lumenal aspect of the granule membrane from contact with secretory enzymes contained within the granules and facilitate the specific release of secretory enzymes during exocytosis at the apical plasma membrane.

The enzymes and the acidic milieu contained in the ZG are released into the lumen of the pancreas through exocytosis by acinar cells. The pH at the apical plasma membrane of the acinar cells, and of the pancreatic lumen in general, is maintained at an essentially neutral or alkaline pH by the fluid and bicarbonate secreted by pancreatic ductal cells. The increased pH at the apical plasma membrane (relative to the acidic pH within the ZG) optimizes the conditions for enzymatic cleavage of the GPI anchor of GP2, resulting in release of GP2 and GP2/PG complexes from the apical membrane. (Scheele et al. (1994) Pancreas 9:139-149, the disclosure of which is incorporated herein by reference in its entirety). The form of GP2 produced by GPI-anchor cleavage is termed globular GP2 (gGP2).

It is believed that the protein of SEQ ID NO:250 is a GP2 protein, and is thus likely involved in regulated membrane trafficking along apical secretory processes in a variety of epithelial cells.

Accordingly, the present invention includes the use of the protein of SEQ ID NO:250, fragments comprising at least 5, 8, 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, or 200 consecutive amino acids thereof, or fragments having a desired biological activity in the modulation of membrane sorting within the trans-Golgi network, assembly of zymogen granule membranes, inactivation of zymgogen granule membranes during granule storage, regulation of zymogen granule membrane trafficking at the apical plasma membrane, release of secretory enzymes during exocytosis at the apical plasma membrane. In such embodiments, the protein of SEQ ID NO:250, or a fragment thereof, is administered to an individual in whom it is desired to increase or decrease any of the foregoing activities. The protein of SEQ ID NO:250 or fragment thereof may be administered directly to the individual or, alternatively, a nucleic acid encoding the protein of SEQ ID NO:250 or a fragment thereof may be administered to the individual. Alternatively, an agent which increases the activity of the protein of SEQ ID NO:250 may be administered to the individual. Such agents may be identified by contacting the protein of SEQ ID NO:250 or a cell or preparation containing the protein of SEQ ID NO:250 with a test agent and assaying whether the

test agent increases the activity of the protein. For example, the test agent may be a chemical compound or a polypeptide or peptide.

Alternatively, the activity of the protein of SEQ ID NO:250 may be decreased by administering an agent which interferes with such activity to an individual. Agents which interfere with the activity of the protein of SEQ ID NO:250 may be identified by contacting the protein of SEQ ID NO:250 or a cell or preparation containing the protein of SEQ ID NO:250 with a test agent and assaying whether the test agent decreases the activity of the protein. For example, the agent may be a chemical compound, a polypeptide or peptide, an antibody, or a nucleic acid such as an antisense nucleic acid or a triple helix-forming nucleic acid.

In one embodiment, the invention relates to methods and compositions using the protein of the invention or part thereof as a marker protein to selectively identify tissues, preferably pancreas and prostate, or to distinguish between two or more possible sources of a tissue sample on the basis of the level of the protein of SEQ ID NO:250 in the sample. For example, the protein of SEQ ID NO:250 or fragments thereof may be used to generate antibodies using any techniques known to those skilled in the art, including those described therein. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, for example, forensic samples,differentiated tumor tissue that has metastasized to foreign bodily sites, or to differentiate different tissue types in a tissue cross-section using immunochemistry. In such methods a tissue sample is contacted with the antibody, which may be detectably labeled, under conditions which facilitate antibody binding. The level of antibody binding to the test sample is measured and compared to the level of binding to control cells from pancreas or prostate or tissues other than pancrease or prostate to determine whether the test sample is from pancreas or prostate. Alternatively, the level of the protein of SEQ ID NO:250 in a test sample may be measured by determining the level of RNA encoding the protein of SEQ ID NO:250 in the test sample. RNA levels may be measured using nucleic acid arrays or using techniques such as in situ hybridization, Northern blots, dot blots or other technques familiar to those skilled in the art. If desired, an amplification reaction, such as a PCR reaction, may be performed on the nucleic acid sample prior to analysis. The level of RNA in the test sample is compared to RNA levels in control cells from pancreas or prostate or tissues other than pancreas or prostate to determine whether the test sample is from pancreas or prostate.

In another embodiment, antibodies to the protein of the invention or part thereof may be used for detection, enrichment, or purification of membranes or zymogen granules using any techniques known to those skilled in the art. For example, an antibody against the protein of SEQ ID NO:250 or a fragment thereof may be fixed to a solid support, such as a chromatograpy matrix. A prepartation containing membranes or zymogen granules is placed in contact with the antibody under conditions which facilitate binding to the antibody. The support is washed and then the membranes or zymogen granules are released from the support by contacting the support with agents which cause the membranes or zymogen granules to dissociate from the antibody.

In another embodiment of the present invention, the protein of SEQ ID NO:250 or a fragment thereof thereof may be used to diagnose disorders associated with altered intercellular communication or secretion. In such techniques, the level of the protein of SEQ ID NO:250 in a patient is measured using techniques such as those described herein. The level of the protein of SEQ ID NO:250 in the patient is compared to the level in control individuals. An elevated level or decreased level of the protein of SEQ ID NO:250 relative to control individuals suggests that the patient is suffering from a defect in intercellular communication or secretion.

In another embodiment, the protein of SEQ ID NO:250 or a fragment thereof is used to facilitate or decrease exocytosis. For example, the protein of SEQ ID NO:250 or fragment thereof may be used to increase or decrease the release of secretory enzymes within pancreatic acinar cells or prostatic cells. Accordingly, the protein of the invention or part thereof may be used to diagnose, treat and/or prevent disorders associated with abnormal membrane trafficking including but not limited to viral or other infections, traumatic tissue damage, and hereditary diseases such as pancreatitis or prostatitis, invasive carcinomas and lymphomas. In such methods, the protein of SEQ ID NO:250, a fragment of the protein of SEQ ID NO:250 , or an agent which increases or decreases the activity of the protein of SEQ ID NO:250 is administered to an individual using techniques such as those described herein.

In another embodiment, the invention relates to methods of using the protein of SEQ ID NO:250 or a fragment thereof in the diagnosis of pancreatitis or prostatitis by detecting an elevation in the level of the protein of SEQ ID NO:250, in a sample of bodily fluid, such as human blood, serum, or urine. The protein may be detected using any method known to those skilled in the art, including those described herein. In some embodiments, the protein of SEQ ID NO:250 or fragment thereof may be detected using the methods described in U.S. Patent Nos. 5436169 or 5663315, the disclosures of which are incorporated herein by reference in their entireties.

References :

U.S. Patent Nos. 5,436,169; 5,663,315
Nucleic Acids Research 18(9):5900, (1990)
Proc. Natl. Acad. Sci. USA 88(7):2898-2902 (1991)
Proc. Natl. Acad. Sci. USA 89:1189-1193 (1992)
Eur. J. Cell Biol. 61:229-238 (1993
Freedman et al., Annals N.Y. Acad. Sci. 713:199-206, 1994.
Scheele et al., Pancreas 9(2):139-149, 1994.

Protein of Seq Id No: 274 (internal designation: 145-56-3-0-D5-CS)

The protein of SEQ ID No: 274 encoded by the cDNA of SEQ ID No: 33 is homologous to the human RNA 3'-terminal phosphate cyclase-like protein 1 (*Rcl1*) (trEMBL accession number CAB89811) which is abundant in the nucleolus.

The RNA 3'-terminal phosphate cyclase, an enzyme originally identified in extracts from human HeLa cells and *Xenopus* oocyte nuclei, catalyzes the ATP-dependent conversion of the 3'-terminal phosphate group into a 2',3'-cyclic phosphodiester at the 3'-end of RNA, resulting in the

activation of the 3' end of RNA molecules. Database searches showed that genes encoding proteins similar to human and *E.coli* human RNA 3'-terminal phosphate cyclase are conserved among eukarya, bacteria and archaea, arguing for an essential function of the enzyme in RNA metabolism (Genschik P. et al. – EMBO J – 1998, 16, p.2955-2967). Similarly analysis of the human RNA 3'-terminal phosphate cyclases and related proteins from other organisms, indicated that they can be divided into 2 subfamilies referred to as RNA 3'-terminal phosphate cyclases (*Rtc*) and RNA 3'-terminal phosphate cyclase-like protein (*Rcl*). These 2 subfamilies share several sequence elements, including a nearly universally conserved amino acid sequence RGxxPxGGGx@ (where x stands for any, and @ for hydrophobic amino acids), designated originally as the cyclase signature, which corresponds to the Prosite signature, although structurally slightly different these 2 subfamilies of proteins have the same function and are involved in RNA metabolism. The cyclase signature is present in the protein of the invention (positions 157 to 167). In addition, this protein also displays othere characteristic signatures of RNA 3'-terminal phosphate cyclase proteins (pfam signature from positions 1 to 368 and eMotif signatures from positions 12 to 44 and from positions 157 to 168).

3'-terminal phosphate cyclases (*Rtc* and *Rcl*) catalyze the conversion of 3'-terminal phosphate to a 2',3'-cyclic phosphodiester in a reaction dependent on ATP, other nucleoside triphosphates being much less active co-factors. With both enzymes, the cyclization of the 3'-phosphate at the 3'-end of RNA occurs by a three-step mechanism as follows : (a) adenylation of the enzyme by ATP; (b) the enzyme acts on RNA-N3' P to produce RNA-N3'PP5'A; (c) a non catalytic nucleophilic attack by the adjacent 2'hydroxyl on the phosphorus in the diester linkage to produce the cyclic end product.

RNA 3'-terminal phosphate cyclase proteins are involved in RNA processing. It has been demonstrated that several eukaryotic and prokaryotic RNA ligases require 2',3'-cylic phosphate RNA ends which suggests that the enzyme is involved in generation or maintenance of cyclic termini in RNA ligation substrates, known to be required by several RNA ligases in both eukaryotes and prokaryotes. These ligases include 2 tRNA-splicing ligases, and the prokaryotic RNA ligase of unknown function that joins RNA ends via atypical 2',5'-phospodiester (Arn E. et al. – RNA structure and Function - Cold spring Harbor Laboratory Press – 1998 p.695-726). The involvement of these ligases in nuclear pre-tRNA splicing is well documented (Zillmann et al. - -Mol Cell Biol – 1991, 11, p5410-5416)(Phizicky E. et al. – J Biol Chem, 1992, 267, p4577-4582) but these enzymes might also function in the ligation of virusoids and viroids (Branch A. et al. – Science – 1982, 217, p1147-1149) (Kibertis et al. – EMBO J – 1985, 4, p817-827)

Alternatively, the cyclase could be responsible for producing cyclic phosphate 3'-ends identified in the spliceosomal U6 small nuclear RNA and some other small RNAs. Furthermore in yeast *Rcl* is associated to U3 small nucleolar RNP (U3 snoRNP) a central component of the 18S ribosomal RNA (rRNA) processing machinery in yeast and vertebrates (Billy E. et al. – EMBO J –

2000, 19, p2115-2126). However it seems that *Rcl* are not a structural component of U3 snoRNP and its association with U3 snoRNP occurs, most probably, in large macromolecular complexes representing nascent ribosmes. In yeast, depletion or inactivation of *Rcl* causes a defect in 18S mRNA synthesis, which leads to a decreased levels of 40S ribosomal sub-units, resulting in an accumulation of free 60S ribosomes and a fall in the amount of polysomes. In Yeast 18S, 5.8S and 25S rRNAs are derived from a long 35S precursor. This 35S pre-rRNA is normally cleaved at the A0 site, yielding 33S pre-rRNA. 33S rRNA is then processed rapidly at sites A1 and A2 to generate 20S pre-rRNA, which is further processed into mature 18S rRNA. Deletion or inactivation of *Rcl* leads to inhibition of processing at sites A0, A1 and A2 (Billy E. et al. – EMBO J – 2000, 19, p2115-2126).

It is believed that the protein of SEQ ID No: 274 or part thereof is involved in RNA processing, probably as a RNA 3'-terminal phosphate cyclase. Preferred polypeptides of the invention are polypeptides comprising amino acids 157 to 167, 1 to 368, 12 to 44 and 157 to 168. Other preferred polypeptides of the invention are fragments of SEQ ID No: 274 having any of the biological activities described herein. Assays of cyclase activity can be carried out using the Norit method as described in the article by Filipowicz (Filipowicz W et al.- Methods Enzymol. – 1990, 181, p.499-510), which disclosure is hereby incorporated by reference in its entity, or any other techniques known to those skilled in the art.

Thus, an embodiment of the present invention relates to compositions and methods of using the protein of the invention or part thereof in in vitro RNA manipulation to isolate small nucleolar RNPs especially, but not limited to U3 snoRNP from biological samples, using immunoprecipitation techniques (Billy E. et al. – EMBO J – 2000, 19, p2115-2126), which disclosure is hereby incorporated by reference in its entity, or any other techniques known to those skilled in the art.

In another embodiment, the protein of the invention or part thereof is used to develop antagonists of the protein of the invention or part thereof in order to inhibit or decrease cellular proliferation. This can be explained by the fact that protein of the invention or part thereof is probably involved in rRNA maturation, thus the use of products that inhibit rRNA maturation prevents the formation of functional ribosomes, which leads to an inhibition of protein synthesis. Cells that are unable to synthesis proteins stop to grow and ultimately die due to the fact that they are unable to regenerate proteins. One preferred embodiment of the invention pertains to the use of the protein of the invention or part thereof to develop these antagonists, which are added to samples or materials as a "cocktail" in association with other antimicrobial substances to stop and/or prevent proliferation of undesired contaminants. For example the protein of the invention or part thereof may be used to inhibit the proliferation of undesired bacteria and or viruses in *in vitro* cultures. In another preferred embodiment of the invention the protein or part thereof could be used to develop antagonists that could be administered to patients suffering from viral and or bacterial infection

particularly viral infections by viruses such as HIV and HCV. This could for example be accomplished by targeting the antagonists to cells infected by the virus or directly to bacteria. Once inside these cells the antagonist will inhibit or at least decrease protein synthesis resulting in an inhibition or a decrease in bacterial and/or viral replication. In yet another preferred embodiment of the invention the protein or part thereof could be used to develop antagonists that could be administered to patients in order to inhibit abnormal and/or unregulated cellular proliferation found in diseases such as cancers, psoriasis, Systemic lupus erythematosus (SLE), arthritis, endometriosis, enteropathy in immunodeficiency virus infection, venous eczema (inducing connective tissue sclerosis in lipodermatosclerosis and causing the reduced reepithelialization tendency in venous ulcers), chronic irritant contact dermatitis (CICD), adult polycystic kidney disease (APKD), ichthyosis, cholesteatoma.

Protein of SEQ ID NOs:303 (internal designation number 187-31-0-0-F12-CS) and 275 (internal designation number 145-59-2-0-A7-CS)

The 148-amino-acid long protein of SEQ ID NO:303, encoded by the cDNA of SEQ ID NO:62, found in fetal kidney and highly expressed in this organ, is homologous to the human RNA-associated protein HSCP250 (SPTREMBLNEW SPTREMBL SWISSPROT accession number AAF36170 and GENESEQP accession number Y84433). In addition, this protein displays significant homology to the ribosomal L27 protein of *D. melanogaster* (GENPEPT GENPEPTNEW accession number AE003576) and to the 50S ribosomal L27 protein of *E.coli* (SWISSPROT accession number P02427). The protein of SEQ ID NO:303 has a putative signal peptide, from amino acid position 13 to 27. According to the PFAM program, the protein of the invention also presents a ribosomal L27 protein signature in position 31 to 81. Amino acid residues in position 64 to 78 are highly similar to the consensus pattern: G-X-[LIVM](2)-X-R-Q-R-G-X(5)-G, where X is any amino acid (the motif found in the protein of SEQ ID NO:303 is G-X-I-I-X-T-Q-R-H-X(5)-G). Potential phosphorylation sites exist in positions 32, 38, 47 (S amino residues), 60 (Y amino residue), 69 and 141 (T amino residues). One of them, the T residue in position 69, is embedded in the ribosomal L27 protein signature described above.

The protein of SEQ ID NO:275, encoded by the cDNA of SEQ ID NO:34, is a 94-amino-acid long variant of the SEQ ID NO:303 protein. While the first 81 amino acid residues of protein of SEQ ID NO:275 are strictly homologous to the first 81 amino acid residues of protein of SEQ ID NO:303, the 13 subsequent amino-acids are different. In addition to the putative signal peptide (position 13 to 27), the ribosomal L27 protein signature (position 64 to 78), and phosphorylation sites (positions 32, 38, 47, 60 and 69), the protein of SEQ ID NO:275 also displays a candidate membrane-spanning segment in position 74 to 94.

Ribosomal protein L27 is one of the proteins of the large ribosomal subunit. L27 belongs to a family of ribosomal proteins which, on the basis of sequence similarities, includes: eubacterial

L27, plant chloroplast L27 (nuclear-encoded), algal chloroplast L27 and yeast mitochondrial YmL2 (gene MRPL2 or MRP7). Among the different ribosomal L27 proteins characterized so far, the one of *E.coli* is probably the best studied. Protein L27 is one of the smallest and the most basic polypeptides in *E.coli* ribosome. Techniques like the measurement of protein exposure by hot tritium bombardment have shown that L27 of the large subunit is well exposed on the surface of the *E.coli* 70S ribosome (Agafonov *et al.*, Proc. Natl. Acad. Sci. 94:12892-12897 (1997)). Chemical and UV-crosslinking studies have demonstrated that L27 is closely associated with domain V of the 23S rRNA, a region that comprises part of the peptidyl transferase center (Osswald *et al.*, Nucleic Acids Res. 18:6755-6760 (1990)). Direct evidence for the presence of L27 at the peptidyl transferase center was obtained through the use of derivatives of $tRNA^{Phe}$ containing photoreactive azidonucleotides within the 3'-terminal $ACCA_{OH}$ sequence (Wower et *al.*, Proc. Natl. Acad. Sci. 86:5232-5236 (1989)). Analysis of a mutant *E.coli* strain in which the *rpmA* gene, which encodes L27, was replaced by a Kanamycin marker, has suggested that L27 contributes to peptide bond formation by facilitating the proper placement of the acceptor end of the A-site tRNA at the peptidyl transferase center (Wower et al., J. Biol. Chem. 273:19847-19852 (1998)). Further, recent studies conducted by Thiede and collaborators have precisely determined RNA-protein contact sites in the 50S ribosomal subunit of *E.coli* (Thiede et al., Biochem. J. 334:39-42 (1998)), showing that Lys-71 and Lys-74 of L27 interact with U-2334 of the 23S rRNA.

It is believed that the proteins of SEQ ID NOs:303 and 275 are human RNA-associated proteins. Preferred polypeptides of SEQ ID NO:303 are polypeptides comprising the amino acids from positions 13 to 27, 64 to 78 and amino acid residues in positions 32, 38, 47, 60, 69 and 141. It is believed that the protein of SEQ ID NO:275 is a 94 amino acid long variant of the 148 amino acid residues protein of SEQ ID NO:303. Preferred polypeptides of SEQ ID NO:275 are polypeptides comprising the amino acids from positions 13 to 27, 64 to 78, 74 to 94 and amino acid residues in positions 32, 38, 47, 60, and 69. Other preferred polypeptides of the invention are fragments of SEQ ID NO:303 or 275 having any of the biological activities described herein.

One embodiment of the present invention involves the use of the present proteins and nucleic acids to specifically identify cells from the kidney, especially from the fetal kidney. Such cells can be detected by virtue of their strong expression of the protein of the invention, and can thus be detected using any standard method for detecting protein expression or activity, including methods involving antibodies, specific nucleic acids, or any other detectable molecule that specifically binds to the polypeptides or polynucleotides of the invention. An ability to specifically detect kidney cells is useful, e.g. for determining the identity of tumor cells as well as for the identification of specific cell types and tissues for, e.g. histological analyses.

In another embodiment of the present invention, the present proteins are used as a component of in vitro eukaryotic translation systems. Such systems represent a widely used tool for protein production with many academic and industrial applications. Similarly, inhibitors of the

protein of the invention, e.g. antibodies or dominant negative forms of the protein, can be used to inhibit in vitro translation systems, e.g. to specifically stop a translation reaction involving a eukaryotic cell extract.

In another embodiment, the proteins of SEQ ID NO:303 or 275 can be used to bind to nucleic acids, preferably RNA, alone or in combination with other substances. For example, the proteins of the invention or part thereof can be added to a sample containing RNAs in optimum conditions for binding, and allowed to bind to RNAs. In a preferred such embodiment, the proteins of the invention or part thereof may be used to purify mRNAs, for example to specifically isolate RNA, e.g. from a specific cell type or from cells grown under particular conditions. Such RNAs could then be reverse transcribed and cloned, could be analyzed for relative expression analyses, etc. In addition, such methods may be used to specifically remove RNA from a sample, for example during the purification of DNA. To carry out any of these methods, the proteins of the invention or part thereof may be bound to a chromatographic support, either alone or in combination with other RNA binding proteins, to form an affinity chromatography column. A sample containing a mixture of nucleic acids to purify is then run through the column. Immobilizing the proteins of the invention or part thereof on a support is particularly advantageous for embodiments in which the method is to be practiced on a commercial scale. This immobilization facilitates the removal of RNAs from the batch of resin-coupled protein after binding, and allows subsequent re-use of the protein. Immobilization of the proteins of the invention or part thereof can be accomplished, for example, by inserting any matrix binding domain in the protein according to methods known to those skilled in the art. The resulting fusion product including the proteins of the invention or part thereof is then covalently, or by any other means, bound to a protein, carbohydrate or matrix (such as gold, "Sephadex" particles, polymeric surfaces).

Still another embodiment of the invention relates to methods of preparing antibodies directed against the proteins of the invention or part thereof. Such antibodies may be used, e.g., in co-immunoprecipitation experiments to separate and purify RNAs associated with the proteins of the invention. To accomplish this, in a sample containing a mixture of nucleic acids, antibodies directed against the protein of the invention may be added in association with protein A or protein G sepharose beads. Immunoprecipitation conditions are well known to those skilled in the art.

The invention further relates to methods and compositions used to modify the proteins of the invention. In a preferred embodiment, K amino-acids of the proteins of the invention are substituted for other basic amino-residues (R residues), as some of these K residues seem to be crucial for RNA interactions (Thiede et al., Biochem. J. 334:39-42 (1998)). Conversely, R residues of the proteins of the invention may be substituted for K residues. These substitutions are predicted to change the specificity and/or the affinity of the proteins of the invention for RNA molecules. Another preferred embodiment may be to perform post-translational modifications of the proteins of the invention, notably at the level of the putative phosphorylation sites described above in positions

32, 38, 47, 60, 69 and 141 of SEQ ID NO:303. By adding negative charges to the proteins of the invention, these phosphorylation sites may modulate the affinity of the protein for RNA molecules. Phosphorylation of T residue in position 69 is of great interest, as it is embedded in the ribosomal L27 protein signature.

In another preferred embodiment, the proteins of the invention or part thereof may be used to visualize RNAs, when the polypeptides are linked to an appropriate fusion partner, or is detected by probing with an antibody.

Another embodiment of the present invention relates to methods and compositions using the proteins of the invention, or part thereof, to associate specific mRNAs to the inner face of lipidic bilayers of liposomes in order to further introduce these mRNAs into the cytoplasm of eukaryotic cells. For example, as described above, the protein of the invention of SEQ ID NO:275 displays both a candidate membrane-spanning segment in position 74 to 94 (at its very carboxy-terminal part), and a ribosomal L27 protein signature in position 64 to 78. Moreover SEQ ID NO:275 is an RNA binding protein. Preferably, specific mRNAs are first associated with the protein of the invention and the RNA/protein complex formed in that way is then mixed with liposomes according to methods known to those skilled in the art. These liposomes are added to an *in vitro* culture of eukaryotic cells. In vivo, such a method might treat and/or prevent disorders linked to dysregulation of gene transcription such as cancer and other disorders relating to abnormal cellular differentiation, proliferation, or degeneration.

In another embodiment, the present proteins and nucleic acids can be used to modulate the rate of cell growth in vitro or in vivo. Studies in Drosophila have shown that a decrease in ribosome function results in a significant inhibition of cell growth. Accordingly, compounds that inhibits the expression or function of the proteins of the invention can be used to inhibit the growth rate of cells, and can thus be used, e.g. in the treatment or prevention of diseases or conditions associated with excessive cell growth, such as cancer or inflammatory conditions. Such compounds include, but are not limited to, antibodies, antisense molecules, dominant negative forms of the proteins, and any heterologous compounds that inhibit the expression or the activity of the proteins.

Protein of SEQ ID NO:269 (internal designation 116-115-2-0-F8-CS)

The protein of SEQ ID NO:269, encoded by the cDNA SEQ ID NO:28, shows homology with the mink whale ribonuclease A (Emmens M., et al., Biochem. J. 157:317-323(1976)) a member of the pancreatic ribonuclease family. In addition, the protein of the invention exhibits 2 membrane spanning segments, the first from amino acid positions 1-21, the second from amino acid positions 179-199. The cDNA SEQ ID NO:28 is composed of 3 exons. Exon 1 is encoded by nucleotides 1-225, exon 2 by nucleotides 226-288, and exon 3 by nucleotides 289-597. The protein of the invention is highly expressed in the testis.

Ribonucleases are proteins that catalyze the hydrolysis of phosphodiester bonds in RNA chains. Pancreatic ribonucleases are pyrimidine-specific ribonucleases present in high quantity in the pancreas of a number of mammalian taxa and of a few reptiles. In addition to their function in the hydrolysis of RNA, ribonucleases have evolved to support a variety of other physiological activities. Such activities include anti-parasite, anti-bacterial, anti-virus, and anti-neoplastic activities, as well as, in some cases, promoting neurotoxicity and angiogenesis. For example, bovine seminal ribonuclease is anti-neoplastic (Laceetti, P. et al. (1992) Cancer Res. 52: 4582-4586), and some frog ribonucleases display both anti-viral and anti-neoplastic activity (Youle, R. J. et al. (1994) Proc. Natl. Acad. Sci. USA 91: 6012-6016; Mikulski, S. M. et al. (1990) J. Natl. Cancer Inst. 82: 151-152; and Wu, Y. -N. et al. (1993) J. Biol. Chem. 268: 10686-10693). In addition, angiogenin is a tRNA-specific ribonuclease which binds actin on the surface of endothelial cells for endocytosis and is then translocated to the nucleus where it promotes endothelial invasiveness required for blood vessel formation (Moroianu, J. and Riordan, J. F. (1994) Proc. Natl. Acad. Sci. USA 91: 1217-1221). Further, eosinophil-derived neurotoxin (EDN) and eosinophil cationic protein (ECP) are related ribonucleases which possess neurotoxicity (Beintema, J. J. et al. (1988) Biochemistry 27: 4530-4538; Ackerman, S. J. (1993) In Makino, S. and Fukuda, T., Eosinophils: Biological and Clinical Aspects. CRC Press, Boca Raton, Fla., pp 33-74). ECP also exhibits cytotoxic, anti-parasitic, and anti-bacterial activities. Finally, an EDN-related ribonuclease, RNase k6, is expressed in normal human monocytes and neutrophils, suggesting a role for this ribonuclease in host defense (Rosenberg, H. F. and Dyer, K. D. (1996) Nuc. Acid. Res. 24: 3507-3513).

It is believed that the protein of SEQ ID NO:269 is a ribonuclease, and is thus capable of hydrolyzing ribonucleic acids, and is involved in the a number of processes including defense against infection and neoplasia, as well as in neurotoxicity and angiogenesis. Preferred polypeptides of the invention are any fragments of SEQ ID NO:269 having any of the biological activities described herein. The ribonuclease activity of the protein of the invention or part thereof may be assayed using any assay known to those skilled in the art, including those described in US patent 5,866,119.

In one embodiment, the present polynucleotides and polypeptides are used to specifically detect testis tissue and cells derived from the testis, as the present protein is overexpressed in this tissue. For example, the protein of the invention or part thereof may be used to synthesize specific antibodies using any technique known to those skilled in the art. Such tissue-specific antibodies may then be used to identify tissues of unknown origin, such as in forensic samples, differentiated tumor tissue that has metastasized to foreign bodily sites, etc., or to differentiate different tissue types in a tissue cross-section using immunochemistry.

The present invention relates to methods and compositions using the protein of the invention or part thereof to hydrolyze one or several substrates, preferably nucleic acids, more

preferably RNA, alone or in combination with other substances. For example, the protein of the invention or part thereof is added to a sample containing a substrate(s) in conditions amenable to enzyme activity, and the protein thus catalyzes the hydrolysis of the substrate(s).

In a preferred embodiment, the protein of the invention or part thereof may be used to remove contaminating RNA in a biological sample, alone or in combination with other nucleases. In a more preferred embodiment, the protein of the invention or part thereof is used to remove contaminating RNA from DNA preparations, to remove RNA templates prior to second strand synthesis and prior to analysis of in vitro translation products. In one such embodiment, the protein of the invention or part thereof is added to a biological sample as a "cocktail" along with other nucleases. The advantage of using a cocktail of hydrolytic enzymes is that one is able to hydrolyze a wide range of substrates without knowing the specificity of any of the enzymes, or even the identity of all of the substrates. Such cocktails of nucleases are commonly used in molecular biology assays, for example to remove unbound RNA in RNAse protection assays. Using a cocktail of hydrolytic enzymes also protects a sample from a wide range of future unknown RNA contaminants from a vast number of sources. For example, the protein of the invention or part thereof is added to samples where contaminating substrates are undesirable. Alternatively, the protein of the invention or part thereof may be bound to a chromatographic support, either alone or in combination with other hydrolytic enzymes, using techniques well known in the art, to form an affinity chromatography column. A sample containing the undesirable substrate is run through the column to remove the substrate. Immobilizing the protein of the invention or part thereof on a support is particularly advantageous for those embodiments in which the method is to be practiced on a commercial scale. This immobilization facilitates the removal of the enzyme from the batch of product and subsequent reuse of the enzyme. Immobilization of the protein of the invention or part thereof can be accomplished, for example, by inserting a cellulose-binding domain in the protein. One of skill in the art will understand that other methods of immobilization could also be used and are described in the available literature. Alternatively, the same methods may be used to identify new substrates.

In another embodiment, the protein of the invention or part thereof may be used to decontaminate or disinfect samples infected by undesirable parasite, bacteria and/or viruses using any of the methods known to those skilled in the art including those described in Youle et al, (1994), supra; Mikulski et al (1990) supra, Wu et al (1993). In a preferred embodiment, the protein is used to eliminate RNA viruses from a sample or in a patient.

In another embodiment, the present invention relates to compositions and methods using the protein of the invention or part thereof to selectively kill cells. The protein of the invention or part thereof is linked to a recognition moiety capable of binding to a chosen cell, such as lectins, receptors or antibodies, thereby generating cell-specific cytotoxic reagents as described in US Patent No. 5,955,073, the disclosure of which is herein incorporated in its entirety.

In another embodiment, the protein of the invention or part thereof is used in the diagnosis, prevention and/or treatment of neoplastic disorders. In one such embodiment, cancer can be treated or prevented in a patient by increasing the activity of the present protein in the patient, particularly within neoplastic or hyperplastic cells within the patient. For example, a polynucleotide encoding the protein of the invention or part thereof, a polynucleotide encoding the protein, or a compound that causes an increase in the expression or activity of the protein, can be administered to the patient, or to cells derived from the patient, in vivo or ex vivo. Preferably, the protein, polynucleotide, or compound is specifically targeted to the neoplastic or hyperplastic cells, for example by intratumoral injection of the molecule or by linking the molecule to a targeting moiety, such as a tumor cell-specific antibody.

In another embodiment, cancer can be treated or prevented in a patient by inhibiting the expression or activity of the protein of the invention in endothelial cells of the patient, in particular within endothelial cells involved in angiogenesis. Such expression or activity can be inhibited in any of a number of ways, for example using antibodies, antisense sequences, ribozymes, dominant negative forms of the protein, as well as small molecule inhibitors of protein activity or expression.

In another embodiment, the present polynucleotide and polypeptide sequences are used to diagnose cancer in a patient. In a typical such embodiment, a biological sample is obtained from a patient, and the level of the present polypeptides or polynucleotides is detected and compared with a control level, where a difference between the level observed in the patient and the control level indicates the presence of cancer in the patient.

In another embodiment, the present protein is is inhibited within cells of a mammal in order to protect cells of the mammal against RNAse-associated neurotoxicity. In a typical such embodiment, the level of the protein is detected within the cells of the patient, where an elevated level of the protein, particularly within neurons, indicates a risk for neurotoxicity. The level of the expression or activity of the protein is subsequently inhibited using any standard method, such as antibodies, antisense molecules, ribozymes, dominant negative forms of the protein, or any other compounds that inhibit the expression or activity of the protein. Preferably, such inhibitors are specifically directed to the neurons of the mammal.

Protein of SEQ ID NO: 390 (internal designation 116-118-4-0-A8-CS)

The present inventors have provided a new gene and protein described in SEQ ID No 149 and 390 respectively, belonging to the carbonic anhydrase (CA) family, more particularly the alpha-CA family. This novel alpha-CA related gene is located on the human chromosome 17q24 region.

The Carbonic anhydrases (EC 4.2.1.1) (CA) are zinc metalloenzymes which catalyze the reversible hydration of carbon dioxide. Nine different active Alpha-carbonic anhydrases (alpha-CA) that catalyze the hydration reaction have been found, as well as at least two alpha-CA-related enzymes. All known carbonic anhydrases from the animal kingdom are alpha-CAs, as opposed to

beta- and gamma-CAs, which are also zinc containing enzymes but are unrelated by sequence. The protein of SEQ ID No. 390 displays significant homology to the pfam Carbonic anhydrase domains amino acids between 20- to 59 of the protein, in particular the motif Gly-Ser-Glu-His in position 45 to 48 of the protein which has been found to be highly conserved in a multi-alignment published by Lovejoy et al. 1998 (Genomics 54, 484-493). The chromosomal localization was found by BLAST alignment with a sequence mapping to chromosome 17 (genbank genomic fragment with accession number AC002090) and another alignment with a genomic fragment (accession number AF064854) which maps the gene in the 17q24 region. The polypeptide of SEQ ID No. 390 displays particularly high homology to a human protein called carbonic anhydrase-related protein 10 (genbank accession number AB036836, published directly in database by Adachi,K. and Nishimori,I).

The human alpha-CAs contain a highly conserved catalytic site which comprises a zinc coordination polyhedron defining an active site located in a large cone-shaped cavity that extends almost to the center of the alpha-CA molecule. One site of the cavity is formed by hydrophobic residues, which the other side contains hydrophillic residues, including Thr199 and Glu106 (referring to CA II enzyme). The zinc ion is located at the bottom of this cleft, and tetrahedrally coordinated to the imidazoles of three histidine residues (His94, His96, His 119, referring to CA II enzyme) and to a water molecule called the 'zinc water' that ionizes to a hydroxide ion with a pK of about 7. (Sly et al., Ann. Rev. Biochem. 64:375-401 (1995). Studies have shown that the Zn-OH-/The199/Glu106 network is important in binding bicarbonate, sulfonamide inhibitors and many anionic inhibitors (Liljas et al., Eur. J. Biochem. 219:1-10).

Improved alpha-CA inhibitors

Of the human alpha-CAs, it has been found that the various isozymes have differing tissue distributions and intracellular localizations. Alpha-CA II for example, is expressed in the cytosol of some cell types in virtually every tissue or organ, while alpha-CA I is expressed in colon and erythrocytes, and alpha-CA IV is expressed on the apical surfaces of epithelial cells of some segments of the nephron, the apical plasma membrane in the lower gastrointestinal tract, and the plasma face of endothelial cells of certain capillary beds. The protein of SEQ ID NO: 390 encoded by the cDNA of SEQ ID NO: 149 has been found by the present inventors to be expressed in testes.

The human alpha-CAs have been found to be involved in a range of important biological functions involving pH regulation, Co_2 and HCO_3- transport, ion transport and water and electrolyte balance. Functions in which alpha-CAs are involved include H+ secretion, HCO_3- reabsorption, HCO_3- secretion, bone resorption, and production of aqueous humor, cerebrospinal fluid, gastric acid and pancreatic juice. Of particular medical interest, CA II has been found to be implicated in osteoporosis, as CA II defects have been found to be a cause of inherited osteoporosis, found along with renal tubular acidosis and brain calcification.

CA activity can be determined by well known means. Assays used to characterize CA isozyme activity are provided for example in Khalifah, J. Biol. Chem. 246:2561-73 (1971); Chen et

al, Biochem 32: 7861-65 (1993); Tu et al., J. Biol. Chem. 258:8867-8871 (1986); and Jewel et al., Biochem. 30:1484-1490 (1991).

Many different inhibitors of CA have been identified, and certain CA inhibitors have been developed as medicaments. CA inhibitors are currently a primary treatment for glaucoma, where inhibition of CA activity reduces intra-ocular pressure by inhibiting formation of aqueous humor. Approved CA inhibitors for glaucoma include Acetazolamide (Diamox®), Methazolamide (Neptazane®), Dorzolamide (Trusopt®) and Brinzolamide (Azopt®).

Improved broad-acting CA inhibitors

In certain treatment settings, there is a need for CA inhibitors capable of inhibiting the broad class of CA isozymes so as to inhibit CO_2 hydration activity. Local (topical) use of CA inhibitors has been found advantageous over systemic application for glaucoma, allowing systemic side effects of CA inhibitors to be avoided. However, current CA inhibitors may have limited efficacy in terms of ability to completely inhibit total CA activity. A topical treatment, Dorzolamide (Trosopt®), for example, is an inhibitor of CA II, but only weakly inhibits CA VI. (Hoyng et al,. Drugs, 50(3): 411-434 (2000). Inhibitors of CA II may be unable to inhibit CA I or other CAs, which is thought to result in decreased drug efficacy because other CAs can compensate for loss of CA II activity (Sly and Hu, Ann. Rev. Biochem. 64:375-401 (1995)). In one example, CA I is five to six times as abundant as CA II in human erythrocytes, but has only about 15% of the activity. Thus, CA I contributes about 50% of the total CA activity (Dodgon et al., J. Appl. Physiol. 64:1492-80 (1988). Moreover, CA I may have different inhibitor sensitivity profile from CA II, as CA I is less sensitive to sulfonamide inhibitors, for example. CA II and CA IV on the other hand, show significant resistance to inhibition with halide ions in comparison to CA I. (Sly et al, (1995), supra) Thus, a significant amount of residual CA activity in a cell or tissue of interest may be due to other CAs, including the polypeptide of SEQ ID No. 390.

Thus, in one aspect, alpha-CA related nucleic acid and polypeptide may be useful for the identification of compounds capable of inhibiting the alpha-CA-catalyzed reversible hydration reaction. In one aspect, the method is carried out to identify or select CA inhibitors capable of inhibiting the activity of the polypeptide of SEQ ID No. 390. In other aspects, the method is carried out to identify or select CA inhibitors capable of broadly inhibiting the activity of a large number of CA enzymes. The nucleic acid and polypeptide sequences of the invention can be used in computer based drug design or for carrying out binding predictions with candidate CA inhibitors in view of the extensive structural information publicly available for CA enzymes. In preferred embodiments, the nucleic acid and polypeptide of the invention is used in drug screening assays. Assays may be cell based or non-cell based assays. In one embodiment, a nucleotide or polypeptide sequence of the invention is brought into contact with a candidate CA inhibitor (such as a CA II inhibitor), and binding of the candidate inhibitor to the polypeptide of the invention, or the activity of the polypeptide of the invention is detected. Activity of the polypeptide of the invention may be CA

activity, or any other suitable activity possessed by the polypeptide of the invention which may be inhibited by binding of the candidate substance. Assays for detecting hydration of carbon dioxide are well known, and referenced above. In preferred embodiments, a panel of CA isozymes including the polypeptide of the invention are screened against the candidate substance, including one or more enzymes selected from the group consisting of CA I, CA III, CA IV, CA VI, and a CA-RP including but not limited to CA-RP VII, CA-RP X and CA-RP XI. In preferred embodiments, a candidate CA inhibitor is selected according to its ability to broadly inhibit CA isozymes capable of catalyzing the hydration of carbon dioxide. Means to conduct such drug screening assays are well known in the art. In one embodiment drug binding is tested, using means described further herein as well as for example in International Patent Publication No. WO 00/58510, the disclosure of which is incorporated herein by reference in its entirety, particularly the section titled "Methods for screening substances interacting with... polypeptides". Drug binding assays on a large panel of isozymes may also be carried out in high throughput format using commercially available binding assay systems (Graffinity Pharmaceutical Design GmbH, Heidelberg, Germany).

The method according to the invention may generally be used to identify or select candidate compounds for the treatment of a disorder characterized by a disorder in pH regulation, Co_2 and HCO_3- transport, ion transport or water and electrolyte balance.

Improved selective CA inhibitors

In other therapeutic strategies, CA inhibitors are delivered orally. However, systemic delivery may affect CA enzymes present in other tissues or organs leading to harmful side effects. It can be expected that CA II inhibitors may also partially or fully inhibit other CA isozymes, such as CA I, or a CA-related protein (CA-RP) such as CA-RP VIII, X, XI or CA RPTP-beta (Tashian et al., In *"Carbonic Anhydrase: New Horizons"*, W.R. Chegwidden, N.D. Carter and Y.H. Edwards, Eds., Birkhauser, Basel); Adachi,K et al., Genbank accession number AB036836; Lovejoy et al. (1998); Peles et al. (1995)) or the CA polypeptide of SEQ ID No 390. CA-RPTP-beta, for example, has a CA domain having no or reduced CA activity but is thought to be involved in the ligand binding or protein complex participation in view of its binding of contactin by an extracellular region. (Peles et al., Cell 82: 251-260 (1995); Tashian et al., (1998), supra). The inhibition of an isozyme such as CA-RPTP-beta or the isozyme of the present invention by systemic treatment using a non-selective drug may result in harmful side effects.

In one example, while oral CA inhibitors for the treatment of glaucoma (eg. acetazolamide) have been effective and without ocular adverse effects, they have shown important systemic effects, including parasthesia of the acra, fatigue, depression, renal stones and gastrointestinal complaints such as nausea and diarrhoea. (Hoyng et al., 2000, supra) Because CA inhibitors are typically used permanently (eg for glaucoma), or over long periods of time, avoiding side effects is particularly important. Selective CA inhibitors capable of inhibiting a CA isozyme of interest to a greater extent than another CA isozyme may thus offer improved means for the treatment of disease.